621.391
Ф 59

Л. М. ФИНК

СИГНАЛЫ ПОМЕХИ ОШИБКИ...

Л. М. ФИНК
СИГНАЛЫ, ПОМЕХИ, ОШИБКИ...

621.391
Ф 59

Л. М. ФИНК

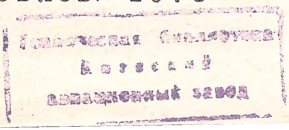
СИГНАЛЫ, ПОМЕХИ, ОШИБКИ...

(ЗАМЕТКИ О НЕКОТОРЫХ НЕОЖИДАННОСТЯХ, ПАРАДОКСАХ И ЗАБЛУЖДЕНИЯХ В ТЕОРИИ СВЯЗИ)

073693
621.391

-2017
ПЕРЕВІРЕНО
К

Москва «Связь» 1978



32.84
Ф59
УДК 621.396

Финк Л. М.

Ф59 Сигналы, помехи, ошибки... (Заметки о некоторых неожиданностях, парадоксах и заблуждениях в теории связи) — М.: Связь, 1978. — 272 с., ил.

55 коп.

Представляет собой заметки различного содержания, относящиеся к статистической теории связи. Наряду с воспоминаниями о задачах, которые приходилось решать автору на протяжении сорока пяти лет его научной работы, здесь разбираются некоторые парадоксы, а также часто встречающиеся ошибки. Основное внимание в книге уделяется не законченным решениям задач, а процессу поиска решения.

Предназначена инженерам и аспирантам, специализирующимся в теории связи и радиотехнике, будет полезна преподавателям и студентам институтов связи, а также широкому кругу читателей, интересующихся историей и практикой науки.

$\Phi \frac{30401-139}{045(01)-78}$ 12—78

ББК 32.84
6Ф0.1

ИБ № 1

ЛЕВ МАТВЕЕВИЧ ФИНК

СИГНАЛЫ, ПОМЕХИ, ОШИБКИ ...

Редакторы В. Л. Черняк, Т. П. Коврова
Художник Е. Н. Алексеев
Художественный редактор А. И. Моисеев
Технический редактор К. Г. Маркоч
Корректор Г. Ф. Кцоева

Сдано в набор 25/V 1977 г. Подп. в печ. 24/VII 1978 г. Т-15202 Формат 70×100¹/₃₂ Гарнитура литературная, печать высокая. Бумага тип. № 2 11,05 усл.-печ. л. 11,06 уч.-изд. л. Тираж 10 000 экз. Изд. № 17288 Зак. тип. 657. Цена 55 к. Издательство «Связь». Москва 101000, Чистопрудный бульвар, д. 2

Отпечатано с набора типографии издательства «Связь», Москва 101000, ул. Кирова, д. 40, в Московской типографии № 4 Союзполиграфпрома Государственного комитета СССР по делам издательств, полиграфии и книжной торговли

г. Москва, И-41, Б. Переяславская, 46

© Издательство «Связь», 1978 г.

ПРЕДИСЛОВИЕ

Предлагаемая читателям книга названа «Сигналы, помехи, ошибки ...», т. е. посвящена основным понятиям теории связи. Но речь в ней идет не столько об ошибках, вызываемых помехами при передаче сигналов, сколько об ошибках, с которыми автору приходилось встречаться в различных статьях, учебниках, диссертациях, выступлениях и т. д. Наряду с ошибками рассматриваются и некоторые мало известные и подчас парадоксальные факты теории связи.

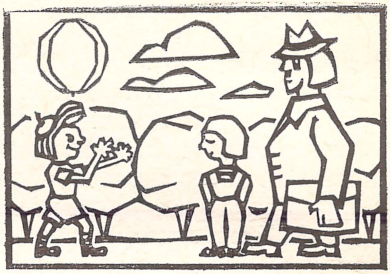
Каждая глава представляет собой самостоятельный очерк и может читаться независимо от других глав, хотя определенная корреляция между главами есть и их последовательность не случайна.

Книга предназначена в основном молодым специалистам, особенно тем, кто собирается посвятить себя научной деятельности. Предполагается, что читатель является специалистом в области связи или радиотехники, знакомым со статистической теорией связи по крайней мере в объеме курса соответствующих вузов. Тем не менее автор пытался написать эту книгу так, чтобы ее чтение было не только полезным, но и в какой-то мере приятным. Насколько ему удалось справиться с этой задачей, смогут судить читатели.

Многие темы были подсказаны автору его товарищами, за что он их горячо благодарит. Особенно большую помощь при обсуждении материалов книги оказали Д. Е. Вакман, Ю. А. Гольдштейн, В. И. Коржик, Б. М. Машковцев, Ю. Б. Окунев, Е. З. Финкельштейн и К. Н. Щелкунов.

Отзывы и замечания следует направлять в издательство «Связь»: 101000, Москва-центр, Чистопрудный бульвар, 2.

1



ОБ ОШИБКАХ, ПАРАДОКСАХ И ЗАДАЧАХ ЭТОЙ КНИГИ

(Вместо введения)

1.1. Об этой книге

Книга эта посвящена в основном различным вопросам теории связи. За последние 15—20 лет в нашей стране и за рубежом появилось множество монографий и тысячи журнальных статей, посвященных этой сравнительно молодой области науки. Зачем же автору понадобилось писать еще одну книгу? Не будут ли в ней еще раз пережевываться давно известные истины, слегка уточняться оценки вероятностей ошибок и пропускных способностей, оттачиваться доказательства старых теорем и т. д.?

Автор надеется, что это не так. Предлагаемая читателю книга как по своему содержанию, так и по форме отличается от большей части написанного (или, по крайней мере, изданного) по теории связи.

Во-первых, она не представляет собой систематического изложения теории связи или какого-нибудь ее раздела. Это сборник отдельных заметок,

связь между которыми скорее ассоциативная, чем логическая. Каждую из них можно читать независимо от остальных.

Во-вторых, в настоящих «серьезных» монографиях и журнальных статьях обычно излагаются окончательные итоги исследований, а поиски, догадки, раздумья, ошибки и их преодоление остаются, как правило, за рамками публикаций. О них знают только сами авторы, да и то с течением времени забывают. В этой же книге, напротив, самое пристальное внимание уделяется не описанию готовых результатов, а процессам их поиска. Именно с этой позиции проводился и отбор материала. Конечно, для систематического курса или для монографии такой подход был бы крайне неэкономичным. Но эта книга и не претендует на то, чтобы заменить систематическое руководство по теории связи, которых написано достаточно много и на любых уровнях сложности и строгости. Она призвана дополнить их путем показа «кухни» научных исследований. Думается, что она должна принести пользу молодым специалистам, особенно тем, которые собираются посвятить себя самостоятельному научному творчеству.

Третья особенность данной книги предопределяется предыдущими и отражена в ее названии. Наибольшее внимание здесь уделяется «подводным камням», встречающимся на пути исследователя. Это различного рода ошибки, которые уже совершались предыдущими поколениями ученых и от которых не гарантированы их последователи. Это также различного рода парадоксальные результаты, заставляющие призадуматься и более тщательно проанализировать изучаемую проблему, а иногда и пересмотреть общепринятую точку зрения.

Содержание предлагаемых заметок довольно разнообразно. Однако они все так или иначе затрагивают основные разделы теории связи, а именно теорию сигналов (в том числе вопросы модуляции), теорию оптимальной обработки сигналов и помехоустойчивости, теорию информации. В частности, возникновение многих парадоксов и ошибок вызвано смешением «мгновенной частоты» сигнала и частоты его спектральных составляющих. Еще больше примеров ошибок дали «изобретатели» всевозможных «вечных двигателей» в связи. Под этим понимаются различные лжепроекты устройств или способов модуляции и обработки сигналов, якобы позволяющие превысить пропускную способность канала или, еще чаще, снизить вероятность ошибки по сравнению с минимально возможной при оптимальном приеме.

Столь же многообразны эти заметки по своему жанру. Одни из них представляют воспоминания автора о решавшихся им отдельных задачах, другие — короткие статьи, посвященные истории связи или какой-либо мало исследованной проблеме, третьи — методические соображения о преподавании, четвертые — отдельные мысли, могущие породить плодотворные идеи. Автор позволил себе в некоторых местах даже нарушить существующую традицию и вести изложение от первого лица.

Таким образом, эта книга не представляет собой «серьезной» монографии. Но в то же время это и не научно-популярная книга, рассчитанная на среднего образованного человека, желающего получить общее представление о неизвестной ему области науки.

1.2. Ошибки

Прежде чем приступить к собственно содержанию книги, следует остановиться на том, что же представляют собой те ошибки и парадоксы, которым посвящена значительная часть последующих заметок. Что такое ошибка — ясно каждому. Это результат, не соответствующий реальной действительности. Причины ошибок в научных исследованиях весьма многочисленны и разнообразны. Конечно, тривиальные ошибки, возникшие в результате обычного недосмотра или описки при расчете, не представляют особого интереса и здесь рассматриваться не будут. Известным литературным примером такой ошибки, повлекшей крах задуманного предприятия, является ошибка героев Жюля Верна, попытавшихся выпрямить земную ось за счет отдачи при выстреле из гигантской пушки. При расчете была допущена описка, в результате которой ожидаемый эффект оказался в 10^6 раз больше действительного.

Значительно полезнее анализ ошибок, вызванных более скрытыми причинами. Такими часто являются не адекватная исследуемому явлению математическая модель, некорректная аппроксимация, смешение сходных, но не тождественных понятий, догматический перенос закономерностей, справедливых для ограниченного круга явлений, на другие и т. п. Говоря об источниках ошибок, нельзя не упомянуть о ряде психологических факторов, способствующих появлению ошибок и мешающих их виновникам осознать их и исправить. Здесь и изящество полученного результата, с которым трудно расстаться, и многообещающие практические применения, которые произвели бы переворот в технике, если бы все было верно, и убеждение ав-

тора в том, что он непризнанный гений. В большинстве случаев все это является результатом низкой общей и научной культуры и недостаточной привычки к самопроверке и самокритике.

Вероятно, еще ни один исследователь (в том числе и автор этой книжки) не смог избежать тех или иных ошибок. Важно, чтобы он не упорствовал в своих заблуждениях, когда они стали для всех очевидными.

Иной раз даже очень крупные ученые ошибались. Так, выдающийся физик Хевисайд, много сделавший для изучения распространения электромагнитных волн, в свое время категорически отрицал возможность создания волноводов. В своей классической работе о теории электромагнетизма он писал следующее (цитирую по книге [8]):

«...Возникает вопрос, можем ли мы пропустить электромагнитную волну вдоль внутренней поверхности трубы наподобие светового луча? Мы, безусловно, можем сделать это при наличии внутри трубы второго провода, потому что это не отличается существенно от случая двух проводов, каждый вне другого. Но это не представляется возможным без внутреннего провода, ибо если мы его уберем, то внутри не останется ничего, на чем могли бы окончиться трубки смещения и вдоль чего они могли бы распространяться...»

Такое высказывание звучит особенно странно, если учесть, что Хевисайд в то время достаточно ясно представлял, что электромагнитные волны могут распространяться в свободном пространстве, причем силовые линии замыкаются сами на себя.

Другой пример того же рода относится к более близкому нам времени. В 50-х годах два очень крупных ученых, много сделавших для развития теории связи, подвергли критическому пересмотру известную формулу Шеннона для пропускной способности C канала с полосой пропускания F и аддитивным гауссовским белым шумом

$$C = F \log_2 (1 + P_c/P_ш), \qquad (1.1)$$

где P_c — мощность сигнала; $P_ш$ — мощность шума.

Вместо этой формулы они получили другую, «уточненную»:

$$C = F [\log_2 (1 + P_c/P_ш) - 1]. \qquad (1.2)$$

Их не смущало то обстоятельство, что вычисленная по их формуле пропускная способность при $P_с = P_ш$ оказывается равной нулю, хотя все знают, что и при равенстве мощностей сигнала и помехи можно худо-бедно какую-то информацию передать. Их даже не удивило, что при переходе от двоичных к другим единицам для количества информации, например к натуральным, т. е. при замене двоичного логарифма натуральным, пропускная способность окажется нулевой при $P_с/P_ш \approx 1{,}71828$. Заметим, что по верной ф-ле (1.1) пропускная способность равна нулю лишь при $P_с/P_ш = 0$ независимо от единиц измерения информации. Замена основания логарифма приводит лишь к соответствующему переходу к другим единицам, а не изменяет характер зависимости C от $P_с/P_ш$, как это происходит с ф-лой (1.2). Лишь через несколько лет авторы ф-лы (1.2) фактически отказались от нее.

Значительное число примеров различных не тривиальных ошибок будет приведено ниже. Существенное внимание будет уделено анализу причин, которые привели к ложному результату.

1.3. Парадоксы

От ошибок следует отличать парадоксы. Термин «парадокс» имеет много значений. В логике парадоксом называется такое рассуждение, которое закономерно приводит к двум противоречивым выводам.

Парадоксами часто называют такие рассуждения, которые приводят к противоречивым результатам вследствие некоторых мало заметных погрешностей в постановке задачи и принятых определениях.

Приведем простой, может быть даже тривиальный, пример, в котором неточность формулировки легко бросается в глаза. Пусть некоторый сигнал передается одновременно по двум параллельным каналам. Каналы между собой статистически неза-

висимы, и вероятность ошибочного приема в каждом из них $p_1 = p_2 = 0{,}1$. Известно, что в одном из этих каналов сигнал принят ошибочно. Какова при этом условии вероятность, что он принят ошибочно в обоих каналах?

Первое решение. В одном канале, по условию, сигнал принят ошибочно. Во втором канале он с вероятностью 0,1 принимается также ошибочно, а с вероятностью 0,9—правильно. Таким образом, при указанном условии вероятность того, что ошибка произошла в обоих каналах, равна 0,1.

Второе решение. При передаче сигнала по двум каналам возможны четыре безусловных исхода:

а) в обоих каналах сигнал принят верно. Вероятность этого исхода равна $p_а = 0{,}9^2 = 0{,}81$;

б) в первом канале сигнал принят верно, а во втором ошибочно. Вероятность этого равна $p_б = 0{,}9 \cdot 0{,}1 = 0{,}09$;

в) в первом канале сигнал принят ошибочно, а во втором верно. Вероятность этого исхода, как и предыдущего, равна $p_в = 0{,}09$;

г) в обоих каналах сигнал принят ошибочно. Вероятность этого равна $p_г = 0{,}1^2 = 0{,}01$.

В соответствии с условием задачи известно, что исход «а» не имел места. Вероятность того, что при этом условии возник исход «г», очевидно,

$$p_г/(p_б + p_в + p_г) = 0{,}01/(0{,}09 + 0{,}09 + 0{,}01) = 1/19.$$

Этот результат почти вдвое меньше, чем полученный в первом решении. Какой же из них верен?

Причиной неоднозначности ответа является некоторая двусмысленность условия задачи. Выражение «в одном из этих каналов сигнал принят ошибочно» можно понимать двояко. Если оно означает «в одном, определенном канале (например, в

том, которому присвоен номер 1), сигнал принят ошибочно», то возможны только два исхода для второго канала и правильным является первое решение. Если же смысл сформулированного условия заключается в том, что «в каком-то одном из двух каналов сигнал принят ошибочно», то верно второе решение.

Различие между этими двумя пониманиями условия можно более наглядно пояснить с помощью формул. Пусть A — событие, заключающееся в том, что ошибка возникла в первом канале, а B — независимое от A (и, следовательно, совместимое[1]) с ним) событие, заключающееся в возникновении ошибки во втором канале. В первом решении ищется условная вероятность $P\ (A \text{ и } B|A)$ или равная ей условная вероятность $P\ (A \text{ и } B|B)$. Она, очевидно, равна $P(A \text{ и } B)/P(A)$ или, учитывая независимость A и B, $P(A)\,P(B)/P(A)=P(B)$. Во втором решении отыскивается другая условная вероятность, а именно $P\ (A \text{ и } B|\bar{A} \text{ или } \bar{B})$. Она равна $P(A \text{ и } B)/[P(A \text{ и } B)+P(\bar{A} \text{ и } B)+P(A \text{ и } \bar{B})]$, где $\bar{A}$ означает событие, противоположное A. Если бы формулировка условия в задаче была более определенной, то никакой двузначности решения не возникало бы.

Заметим, что аналогичные парадоксы, связанные с неточностью формулировки условия, давно известны и упоминают-

---

[1]) Кое-где в литературе встречаются упоминания о событиях «независимых и в то же время несовместимых». Это, конечно, основано на недоразумении, поскольку несовместимость подразумевает весьма «сильную» зависимость. В самом деле, если события A и B независимы, то это значит, что $P(A|B)=P(A)$ и $P(B|A)=P(B)$. Если же они несовместимы, то $P(A|B)=0$ и $P(B|A)=0$. Объединяя эти равенства, получим $P(A)=0$ и $P(B)=0$. Следовательно, независимыми и несовместимыми могут быть одновременно только события, имеющие нулевую вероятность.

ся во многих популярных книгах. Типичный пример такой задачи — у Иванова двое детей, известно, что *один из них* мальчик; какова вероятность того, что у него два мальчика? С другой стороны, у Петрова тоже двое детей, причем известно, что *старший* мальчик; какова вероятность того, что у Петрова два мальчика? Предполагается, конечно, что рождение мальчика или девочки — события независимые и равновероятные. После обсуждения предыдущего примера читатель легко сообразит, что вероятность иметь двух мальчиков равна для Иванова 1/3, а для Петрова — 1/2.

Иногда термин «парадокс», или «парадоксальное решение» применяют и к вполне однозначному и закономерно полученному выводу, если он противоречит тому, что ожидалось получить интуитивно, или, как иногда говорят, противоречит «здравому смыслу». Приведем в заключение пример такой задачи.

Три пеленгатора одновременно пеленгуют некоторый объект. Если бы пеленгование производилось абсолютно точно, то три линии пеленга, проведенные от каждого из пеленгаторов, пересеклись бы в одной точке, в месте нахождения пеленгуемого объекта (предполагается, что объект не лежит на одной прямой ни с одной из пар пеленгаторов). В действительности пеленг получается с некоторой ошибкой, и поэтому линии пеленга пересекаются не в одной точке (рис. 1.1), а образуют некоторый треугольник ABC. Очевидно, что если ошибки пеленгования малы, то объект находится где-то недалеко от точек A, B и C. Лет тридцать тому назад среди специалистов, занимающихся практикой пе-

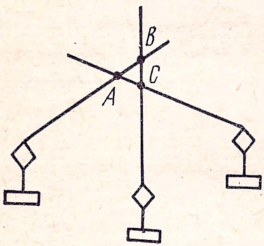


Рис. 1.1. Пеленгование тремя пеленгаторами

ленгования, было очень распространено мнение, что в отсутствии систематических ошибок пеленгуемый объект с большой вероятностью (во всяком случае с вероятностью, большей 1/2) находится внутри треугольника ABC. Проверим это утверждение.

Введем естественное допущение, полагая, что угловая ошибка пеленгования является случайной величиной, принимающей с равной вероятностью положительные и отрицательные значения. Найдем при этом предположении вероятность того, что пеленгуемый объект находится внутри треугольника ABC.

Решим эту задачу, предполагая, что пеленгуемый объект O находится вне треугольника abc, в вершинах которого расположены пеленгаторы. Для определенности выберем обозначения пеленгаторов так, чтобы прямая Ob лежала внутри угла aOc.

Существует восемь равновероятных сочетаний знаков угловых ошибок пеленгования, которые показаны на рис. 1.2. Легко видеть, что только при двух сочетаниях, когда ошибки пеленгаторов a и c имеют одинаковый знак, а ошибка b — противоположный им знак, пеленгуемый объект оказывается внутри треугольника ABC. Не менее легко убедиться в том, что этот результат не зависит от величин ошибок, а только от их знаков. Поскольку все восемь сочетаний знаков ошибок равновероятны, искомая вероятность того, что объект пеленгования находится внутри треугольника ABC, равна $2/8 = 1/4$.

Совершенно аналогично и с тем же результатом рассматривается случай, когда объект пеленгования находится внутри треугольника abc. Различие заключается лишь в том, что при таком располо-

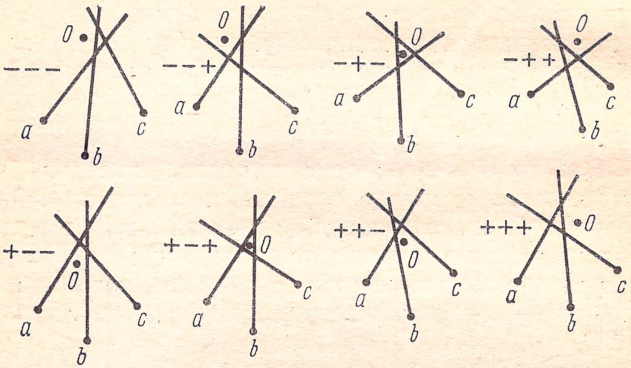


Рис. 1.2. К задаче о трех пеленгаторах

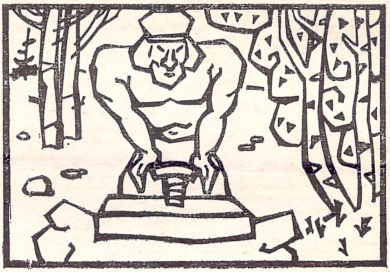
жении благоприятными являются два случая, когда знаки всех трех угловых ошибок одинаковы.

Неожиданность полученного решения состоит не только в том, что найденная вероятность оказалась значительно меньшей, чем это ожидалось. Примечателен тот факт, что эта вероятность не зависит ни от дисперсий угловых ошибок, ни от вида распределения их вероятностей (которые могут быть и не одинаковыми), ни от расположения пеленгаторов. Достаточно, чтобы знаки угловых ошибок для каждого пеленгатора были равновероятны.

Такой результат, на первый взгляд, даже противоречит здравому смыслу. Как может быть, что для очень точных пеленгаторов, у которых дисперсия ошибки ничтожно мала, искомая вероятность такая же, как и для самых грубых? Однако если немного вдуматься, то никакой нелепости в этом не обнаруживается. При точных пеленгаторах треу-

гольник ABC будет, как правило, очень малым, а при грубых пеленгаторах — большим. Поэтому, несмотря на то, что вероятность нахождения объекта внутри этого треугольника в обоих случаях одинакова, точные пеленгаторы позволяют оценить его положения значительно лучше, чем грубые.

Приведенные здесь примеры имеют лишь весьма отдаленное отношение к теории связи. Ниже будут рассмотрены ошибки и парадоксы, характерные для проблем связи. Главы 2—8 посвящены в основном детерминированным сигналам и цепям. Далее наибольшее внимание уделяется статистической теории связи и некоторым смежным с ней вопросам.



2

СПЕКТР И МГНОВЕННАЯ ЧАСТОТА

2.1. Спор о спектре амплитудномодулированного сигнала

Частота и спектр — это, пожалуй, два понятия, которыми больше всего пользуются современные инженеры связи. Трудно поверить, что, например, техника радиосвязи развивалась в течение десятка лет без использования каких-либо представлений о спектрах радиосигналов. А между тем так оно и было. Даже тогда, когда в технику телефонной связи спектр сигнала вошел в качестве одной из основных инженерных характеристик, в области радиосвязи (в те времена главным образом радиотелеграфной) о спектре сигнала почти не говорили. Причину этого понять нетрудно. Спектры применявшихся тогда радиосигналов (пачек затухающих синусоид, подвергнутых амплитудной манипуляции, а позднее — амплитудноманипулированные незатухающие синусоиды при частоте манипуляции, не превышающей нескольких десятков герц) были значительно уже, чем полоса пропускания контуров в передатчиках и приемниках. Поэтому вопрос о

спектре радиосигнала с точки зрения инженерной практики был совершенно неактуальным.

Положение изменилось с появлением радиотелефонии, которая осуществлялась вначале только с помощью амплитудной модуляции. Передаваемый сигнал $s(t)$ при этом может быть представлен в форме

$$s(t) = A(t)\cos(\omega_0 t + \varphi_0), \qquad (2.1)$$

где $A(t)$ — огибающая (или «переменная амплитуда») сигнала; $\omega_0 = 2\pi f_0$ — его круговая частота[1]; φ_0 — его начальная фаза.

Отличие этого сигнала от чисто гармонического $A\cos(\omega_0 t + \varphi_0)$ с постоянной амплитудой A на первый взгляд ничтожное. Все дело только в том, что в одном случае A — постоянная, а в другом — переменная величина. Однако, как хорошо известно, спектры этих сигналов качественно различны — в случае постоянной амплитуды A весь спектр состоит из одной единственной составляющей с частотой ω_0 (монохроматический спектр), в случае переменной A спектр оказывается сложным и характер его определяется видом функции $A(t)$. В частности, при модуляции одним тоном с частотой $\Omega < \omega_0$ сигнал имеет вид

$$A_0[1 + m\cos\Omega t]\cos(\omega_0 t + \varphi_0) = A_0\cos(\omega_0 t + \varphi_0) + \\ + \frac{m}{2}A_0\cos[(\omega_0 + \Omega)t + \varphi_0] + \frac{m}{2}A_0\cos[(\omega_0 - \\ - \Omega)t + \varphi_0], \qquad (2.2)$$

т. е. разлагается на три гармонические составляю-

---

[1] В дальнейшем везде слово «круговая» будет опускаться и величина ω будет называться просто частотой. Использование в (2.1) косинуса, а не синуса не меняет ничего принципиально, но оказывается более удобным в дальнейшем.

073693

щие с частотами ω_0 (несущая частота), $\omega_0+\Omega$ (верхняя боковая частота) и $\omega_0-\Omega$ (нижняя боковая частота).

Все это сейчас отлично известно любому студенту и трудно поверить, что в 20-е и даже в 30-е годы нашего столетия некоторые крупные инженеры, в том числе изобретатель вакуумного диода английский ученый Флеминг [47], возражали против концепции боковых частот. Конечно, это возражение не было направлено против ф-лы (2.2). Никто не собирался опровергать тригонометрическую формулу разложения произведения косинусов. Флеминг [47] не отрицал, что при сложении сигналов от трех точно сфазированных генераторов гармонических колебаний—$A\cos(\omega_0 t+\varphi_0)$, $(m/2)\times$ $\times A_0\cos[(\omega_0+\Omega)t+\varphi_0]$ и $(m/2)A_0\cos[(\omega_0-\Omega)t+$ $+\varphi_0]$ — можно получить в точности такой же сигнал, как и при модуляции гармонического колебания гармоническим же низкочастотным сигналом $\cos\Omega t$ с глубиной модуляции m. Спор шел о том, содержатся ли реально боковые частоты $\omega_0+\Omega$ и $\omega_0-\Omega$ в модулированном сигнале.

Флеминг и его единомышленники полагали, что преобразование (2.2) является одним из многих возможных математических представлений и ничего не говорит о реальном существовании боковых частот. Например, даже простой гармонический сигнал можно всегда разложить различным образом на сумму нескольких других сигналов. Обозначим, например, через $M(t)$ периодическую «треугольную» функцию с периодом $T=2\pi/\omega_0$, определяемую на интервале $-T/2\leqslant t\leqslant T/2$ выражением

$$M(t)=\begin{cases}1+4t/T & \text{при } t<0,\\ 1-4t/T & \text{при } t\geqslant 0\end{cases} \qquad (2.3)$$

и периодически продолженную вне этого интервала, а через $N(t)$ — функцию, равную $\cos\omega_0 t - M(t)$ (рис. 2.1). Тогда по определению

$$\cos\omega_0 t = M(t) + N(t). \qquad (2.4)$$

Но можно ли на основании этого утверждать, что в косинусоиде реально содержится «треугольная» функция $M(t)$?

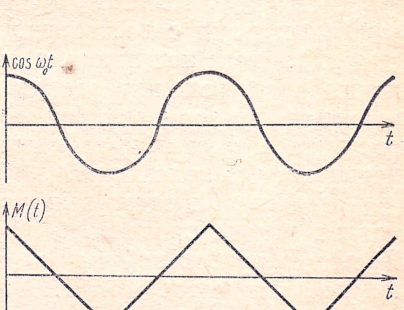


Рис. 2.1. Разложение гармонического сигнала

Что же касается модулированного сигнала (2.1), то мы видим, говорит Флеминг, что его частота равна ω_0. Поэтому никаких других частот в нем на самом деле не содержится, какие бы математические преобразования с ним не производили.

Как видите, доводы Флеминга были далеко не так наивны, как их иногда представляют. Не содержат ли они некоторое рациональное зерно?

К этому вопросу мы еще вернемся. Пока же отметим, что следующий шаг в рассуждениях Флеминга был безусловно ошибочным. Он утверждал, что поскольку в сигнале (2.1) существует только частота ω_0, то его может выделить контур, настро-

енный на эту частоту, и чем острее будет резонансная кривая этого контура, тем лучше он отделит этот сигнал от других сигналов с другими частотами ω_1, ω_2, ..., лишь бы эти частоты отличались от ω_0. Отсюда делался вывод, что, в принципе, можно в заданном диапазоне частот разместить сколько угодно амплитудномодулированных (АМ) сигналов и они не будут мешать друг другу, если их выделять контурами с достаточно высокой добротностью. Поэтому нет никаких оснований к установлению частотных интервалов между полосами, отводимыми различным радиостанциям, и «плотность населения эфира» лимитируется только селективностью приемника.

Ошибочность этого вывода была полностью доказана в ходе протекавшей в 1930 г. дискуссии. Большую роль в ней сыграл известный советский ученый, академик Л. И. Мандельштам. Многие участники дискуссии отмечали, что боковая частота (и вообще любая составляющая спектра) становится реальной, как только используется селективная система, способная ее выделить. Это относится не только к гармонической составляющей. Можно выделить, например, составляющую $M(t)$ (2.3) из косинусоиды (2.4), построив селективную систему в виде параметрического фильтра, для которого треугольная функция $M(t)$ является собственной функцией (подробно см. [8]). Однако можно и не пользуясь представлением о боковых частотах показать, что повышение добротности контуров не позволит разместить без взаимных помех сколько угодно АМ сигналов в заданной полосе частот. Для этого достаточно вспомнить, что чем выше добротность контура, тем больше и его инерционность, и поэтому приходится больше времени затрачивать на любое изменение амплитуды коле-

баний в контуре. При увеличении добротности возникают условия, при которых, например, АМ сигнал (2.2) вызовет в контуре колебания, амплитуда которых за период частоты Ω не будет успевать заметно измениться. Со спектральной точки зрения это значит, что контур практически пропускает только несущую частоту и сильно ослабляет боковые частоты.

Взгляды Л. И. Мандельштама хорошо изложены в книге [13] и поэтому подробно здесь не обсуждаются. Проблема реальности боковых частот уже давно стала достоянием истории. О ней писалось много раз в различных работах. Мы вспомнили здесь об этом не для того, чтобы еще раз доказывать существование боковых частот. Нас интересует первопричина ошибки Флеминга. Но прежде чем говорить о ней, вспомним еще об одной дискуссии, которая велась примерно в те же годы.

2.2. Спор об узкополосной частотной модуляции

В ходе полемики о боковых частотах АМ сигнала практически всем специалистам стало ясно, что всякое изменение амплитуды сигнала приводит к расширению его спектра. При амплитудной модуляции гармоническим сигналом с частотой Ω спектр не может быть уже, чем 2Ω, и поэтому в диапазоне частот шириной Δ нельзя разместить больше чем $\Delta/2\Omega$ АМ сигналов так, чтобы их можно было разделить линейными частотно-селективными цепями. Здесь под Ω нужно понимать максимальную частоту спектра модулирующего сигнала. При радиотелефонии (модуляции речью) обычно $\Omega \approx 2\pi \cdot 3000$ Гц.

А все же расставаться с идеей более эффективного использования диапазона частот не хотелось.

И вот, у некоторых инженеров появилась мысль — заменить амплитудную модуляцию частотной. Изложим эту мысль примерно так, как ее описывали в те далекие времена, когда теории частотной модуляции (ЧМ) еще не существовало.

Итак, пусть исходным модулируемым колебанием будет гармонический сигнал $A\cos\omega_0 t$. Но, в отличие от АМ, будем сохранять амплитуду A постоянной. Вместо нее будем модулировать частоту ω, положив $\omega=\omega_0+\Delta\omega x(t)$, где $x(t)$ — первичный модулирующий сигнал, в простейшем случае $x(t)=\cos\Omega t$. Предполагается, что $x(t)$ — безразмерная функция, принимающая значения в пределах от -1 до $+1$, величина $\Delta\omega$ — девиация частоты, выбираемая, вообще говоря, произвольно.

Поскольку амплитуда A остается постоянной, боковых частот здесь не возникает. Правда, частота такого сигнала изменяется, но только в пределах от $\omega_0-\Delta\omega$ до $\omega_0+\Delta\omega$. Величину же девиации $\Delta\omega$ можно выбрать достаточно малой, например 300 Гц или даже 30 Гц. Таким образом, можно передавать такой сигнал в узкой полосе частот, соответственно равной 600 Гц или 60 Гц, вместо 6000 Гц при АМ, т. е. в 10—100 раз увеличить число каналов, размещенных в заданном диапазоне частот. Прием же таких сигналов осуществляется, в принципе, очень просто — сигнал подается на слегка расстроенный колебательный контур так, чтобы средняя частота ω_0 соответствовала середине того участка резонансной кривой, которую с некоторым приближением можно считать линейной. На выходе такой цепи частотномодулированный сигнал оказывается дополнительно промодулированным по амплитуде (рис. 2.2), после чего обычный амплитудный детектор может восстановить первичный (модулирующий) сигнал. Из рис. 2.2 видно,

что величина девиации никак не должна сказываться на рассматриваемых операциях, если только добротность контура выбрана так, чтобы его полоса пропускания, отсчитываемая на уровне 0,5, была немного больше удвоенной девиации. Поэтому, со-

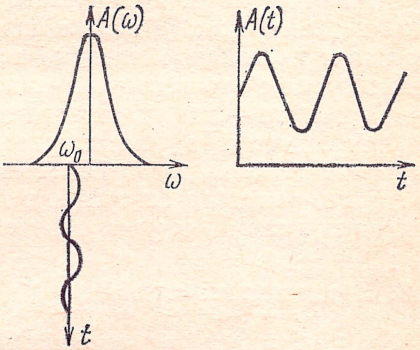


Рис. 2.2. Детектирование ЧМ сигнала на скате резонансной кривой

вершенствуя колебательный контур и добиваясь повышения его добротности, можно уменьшать величину девиации и, следовательно, размещать все большее число каналов в заданном диапазоне частот.

Ошибочность такой точки зрения еще задолго до выступления Флеминга отметил Карсон [17]. Сигнал ЧМ имеет спектр, в состав которого входят боковые частоты вида $\omega_0 \pm k\Omega$, где k — любое натуральное число, и, таким образом, теоретически занимает бесконечно широкую полосу. Если даже ограничиться той полосой частот, в которой расположена только наиболее существенная часть спектра, то и тогда необходимо будет учитывать хотя бы первую пару боковых частот, т. е. частоты $\omega_0 \pm \Omega$. Таким образом, спектр ЧМ сигнала опреде-

ляется не только девиацией и не может быть более узким, чем спектр АМ сигнала при одинаковых модулирующих сигналах. Несмотря на ясность этого вопроса, еще в 1929 г. некий Робинсон получил в США патент на применение узкополосной ЧМ для сужения спектра передаваемого сигнала.

Как известно, в дальнейшем применение ЧМ вело к расширению, а не сужению спектра, что увеличивало помехоустойчивость связи. Но сейчас нас интересует не это обстоятельство. Поставим следующий вопрос — нет ли какого-то общего источника погрешности в рассуждениях одних авторов, отрицавших реальность боковых частот при АМ, и других, пытавшихся сузить используемую полосу частот применением узкополосной ЧМ.

2.3. В чем корень ошибки?

Действительно, такой источник ошибочных рассуждений существует и рассмотрение его может быть весьма полезным. К сожалению, в известной учебной и монографической литературе ему не уделено достаточного внимания. Как будет показано ниже, и в настоящее время отдельные специалисты совершают ошибки, в основе которых лежит неумение четко различать два похожих, но отнюдь не совпадающих понятия — *мгновенная частота* сигнала и *частота* его спектральной составляющей. При спектральном представлении сигнала в виде ряда Фурье он выражается суммой гармонических составляющих, каждая из них является функцией, заданной на всей бесконечной оси времени и характеризуемой амплитудой, частотой и начальной фазой. Строго говоря, понятия «амплитуда» и «частота» можно применять только к такой гармонической функции, как $A\cos(\omega_0 t+\varphi)$, существующей при

$-\infty < t < \infty$. Процессов, описываемых такими функциями, в природе, по-видимому, не существует. Однако, представляя реальный процесс рядом Фурье (или, с некоторыми несущественными оговорками, интегралом Фурье), его разлагают именно по таким бесконечно существующим функциям с неизменными амплитудой, частотой и начальной фазой.

Но спектральное представление сигнала не является единственно возможным и далеко не всегда самым удобным. Часто реальный сигнал записывают в виде

$$s(t) = A(t)\cos[\omega_0 t + \varphi(t)]. \qquad (2.5)$$

По своей форме такая запись очень похожа на выражение гармонического сигнала. Отличие заключается лишь в том, что «амплитуда» $A(t)$ и «начальная фаза» $\varphi(t)$ (в 2.5) зависят от времени. Частный случай такого выражения уже встречался в (2.1). Заметим, что в форме (2.5) можно записать и *финитный* сигнал, т. е. отличный от нуля только на некотором отрезке времени — от t_1 до t_2. Для этого достаточно положить $\varphi(t) = \pi/2 - \omega_0 t$ при $t < t_1$ и $t > t_2$.

Величина ω_0 в (2.5) не зависит от времени и на первый взгляд играет ту же роль, что и частота ω_0 в представлении простого гармонического сигнала:

$$A\cos(\omega_0 t + \varphi_0). \qquad (2.6)$$

Однако удобнее определить «частоту» несколько иначе.

Рассматривая выражение (2.6) и обозначая в нем «полную фазу» (т. е. аргумент косинуса) через $\Phi(t) = \omega_0 t + \varphi_0$, легко видеть, что частота ω_0 равна производной начальной фазы $d\Phi(t)/dt$. Аналогично для сложного сигнала (2.5) обозначим и определим «мгновенную частоту» $\omega(t) = d\Phi/dt =$

$=\omega_0+d\varphi/dt$. Она является функцией времени и в общем случае не равна ω_0.

Автор приносит извинения искушенному читателю за то, что он пересказывает широко известные определения. Однако это оправдано необходимостью подчеркнуть некоторые факты, на которые иногда не обращают внимания.

Итак, сигнал (2.5) характеризуется мгновенной амплитудой (или огибающей) $A(t)$, мгновенной частотой $\omega(t)$ и мгновенной начальной фазой $\varphi(t)$. Но этот же сигнал можно представить в виде суммы спектральных составляющих, каждая из которых характеризуется своей амплитудой, своей частотой и своей начальной фазой. И в том и в другом случаях применяется термин «частота» по исторически сложившейся традиции. Однако свойства мгновенной частоты и частоты спектральной составляющей во многом различны. Одинаковой для них является только размерность (радиан в секунду или после деления на 2π — герц). Что же касается других свойств, то их различие представлено ниже.

Мгновенная частота	*Частота гармонической составляющей спектра*
1. Является функцией времени	Не зависит от времени
2. Для данного сигнала в данный момент времени принимает одно единственное значение	Для данного сигнала в любой момент времени существует конечное, счетное или несчетное множество спектральных составляющих с различными частотами
3. Может изменяться при прохождении сигнала через линейную цепь с постоянными параметрами	При прохождении сигнала через линейную цепь с постоянными параметрами частоты спектральных составляющих не изменяются. Могут измениться только их амплитуды и начальные фазы

4. Не может являться аргументом передаточной функции цепи	Является аргументом передаточной функции линейной цепи

Некоторые другие свойства мгновенной частоты будут рассмотрены в гл. 6.

Обратим внимание на третье свойство спектральных составляющих. Именно оно является причиной столь широкого использования спектрального представления сигналов. Учитывая свойство суперпозиции, характеризующее линейные цепи, можно свести изучение прохождения сложного сигнала через линейную цепь к прохождению отдельных его составляющих. В качестве таких составляющих удобно принять гармонические сигналы, так как после прохождения через линейную цепь с постоянными параметрами они остаются гармоническими и сохраняют свою частоту. Если знать изменения, которым подвергается амплитуда и фаза такого гармонического сигнала при прохождении через цепь (т. е. другими словами, передаточную функцию цепи), то, пользуясь методом суперпозиции, можно сразу получить результат прохождения всего сложного сигнала.

Представление (2.5), в свою очередь, очень удобно при изучении прохождения сигналов через безынерционные нелинейные цепи, например квадратор. При определенных условиях $A(t)$ и $\omega(t)$ оказываются медленно изменяющимися функциями времени, что позволяет получить ряд полезных точных или приближенных результатов. Но для изучения прохождения сигнала через линейную цепь представление (2.5) не всегда удобно.

Заметим также, что для идеального гармонического сигнала $A\cos(\omega_0 t+\varphi_0)$, $-\infty<t<\infty$, мгно-

венная частота ω_0 является постоянной и совпадает с частотой его единственной спектральной составляющей.

После этого краткого отступления легко понять, что явилось первопричиной ошибок в спорах о существовании боковых частот при АМ и о полосе частот, занимаемой узкополосным ЧМ сигналом. Только в результате смешения понятий мгновенной частоты и частоты спектральной составляющей могли возникнуть те недоразумения, о которых здесь идет речь.

Действительно, АМ сигнал (2.1) $s(t)=A(t)\cos(\omega_0 t+\varphi_0)$ имеет переменную огибающую $A(t)$ и постоянную мгновенную частоту ω_0. По аналогии с простым гармоническим сигналом (2.6) Флеминг и его единомышленники отождествляли эту мгновенную частоту с частотой якобы единственной спектральной составляющей. Точно так же изобретатели «узкополосной ЧМ» не видели разницы между множеством частот, входящих в спектр сигнала, и множеством значений, принимаемых мгновенной частотой. Поэтому они отождествляли ширину спектра с областью изменения мгновенной частоты сигнала.

Отметим одно интересное обстоятельство. В те времена, о которых идет речь, радиоинженеры, говоря о частоте негармонического сигнала, обычно имели в виду мгновенную частоту, хотя этот термин тогда еще широко не использовался. Такое представление о частоте для человека, привыкшего оперировать с простыми гармоническими сигналами, проще и естественнее спектральных представлений. Именно на этой почве возникали парадоксы и ошибочные заключения, подобные изложенным выше.

2.4. Обжегшись на молоке...

Время шло, различные виды модуляции внедрялись в технику связи, возникали многообразные задачи о прохождении сложных сигналов через линейные цепи, эти задачи решались чаще всего методом спектрального анализа с использованием передаточной функции цепи. Примитивный подход с подменой спектральной частоты мгновенной приводил к ошибкам, а спектральный анализ, применение амплитудно-частотных и фазо-частотных характеристик становились все более обычными и понятными инженерам. Неудивительно, что в результате этого у ряда специалистов возникло предубеждение против самого понятия «мгновенная частота», которое к тому времени уже было явно сформулировано.

Пожалуй, в наиболее ярком виде это неприятие мгновенной частоты выразил Дж. Шекел в своем «письме в редакцию» журнала «Proceedings of the Institute of Radio Engeneers» [54], в котором говорилось следующее:

«В недавно напечатанной статье систематически применяется термин «мгновенная частота». Многие авторы уже отмечали, что этот термин является ошибочным и вводящим в заблуждение, особенно при рассмотрении частотной модуляции. Ниже мы предполагаем показать, почему этот термин не применим, и надеемся, что после этого он навсегда исчезнет из словаря инженеров связи».

Далее в этом письме говорится, что мгновенную частоту можно определить, совмещая в данной точке заданную функцию и две ее производные с синусоидой, удовлетворяющей, как известно, дифференциальному уравнению

$$f'' + \omega^2 f = 0. \qquad (2.7)$$

Поэтому мгновенную частоту любого сигнала следует вычислять как

$$\omega = \sqrt{-f''/f}. \qquad (2.8)$$

Приведя такое произвольное определение мгновенной частоты, Шекел применяет его к сигналу $f(t) = \sin g(t)$, причем «мгновенная частота» оказывается равной $\sqrt{(g')^2 - g'' \operatorname{ctg} g}$ вместо, как он говорит, ожидаемого «интуитивного» значения $g'(t)$.

Получив явно нелепый результат, Шекел вместо того, чтобы отвергнуть свое определение, ополчается, в первую очередь, на представление частоты как производной фазы и в итоге вообще отвергает понятие мгновенной частоты.

Заканчивая свое письмо, Шекел пишет:

«Поэтому легко понять кажущиеся парадоксы в такого рода утверждениях: «максимальный отклик резонансного *RLC* контура на напряжение с переменной частотой имеет место не тогда, когда мгновенная частота совпадает с резонансной частотой контура» или «спектр ЧМ колебания значительно шире области изменения мгновенной частоты». Эти утверждения ошибочны, поскольку они основаны на «интуитивном», но неверном понимании термина» (см. также [55]).

Обратим внимание на интересный факт. С 1929 г., когда Робинсон получил патент на узкополосную ЧМ, до 1953 г., когда Шекел опубликовал свое письмо, прошло 24 года. За это время отношение к спектру и мгновенной частоте в корне изменилось. Робинсон считал реальными только мгновенные частоты, Шекел — только спектр.

В одном из писем с возражениями Шекелу Хок [52] справедливо заметил, что обобщать ур-ние (2.7) на случай переменной ω также нелепо (improper), как подставлять $\omega(t)$ непосредственно в выражение для гармонического сигнала $A\cos(\omega t+\varphi)$.

Не следует думать, что Шекел был одинок. Он, может быть, только в более категоричной форме высказал мысли, которые в 50-х годах разделяли многие ученые. Приведем в качестве примера одно место из работы академика А. А. Харкевича [49], которая в то время была бесспорно лучшей из книг, трактующей вопросы спектрального анализа сигналов, во многом опережала аналогичные зарубежные работы и сыграла выдающуюся роль в воспитании не одного поколения инженеров связи. В одном из «добавлений» к этой книге автор ее рассматривает широко распространенную в те времена теорию частотномодулированного радиолокационного альтиметра.

Напомним современному читателю сущность этой теории. Альтиметр предназначен для измерения высоты самолета над поверхностью земли. Для этого с самолета посылается сигнал, модулированный по частоте по треугольному закону. Отраженный от земли сигнал принимается бортовым приемником. Его частота изменяется по тому же закону, но с запаздыванием на $\tau=2h/c$, где h — высота, c — скорость электромагнитных волн.

На рис. 2.3 сплошной линией показано изменение частоты излучаемого сигнала, а пунктиром — принимаемого. На протяжении большей части времени разность частот передаваемого и принимаемого сигналов остается постоянной и пропорциональна τ, а следовательно, и h. Сложив излучаемый и отраженный сигналы, получим биения разностной

частотой, которую можно выделить и измерить частотомером, определив тем самым высоту полета.

По этому поводу А. А. Харкевич писал:

«Странно, что это наивное и в корне неверное описание удерживается в течение долгих лет в технической литературе, тогда как общие ошибочные

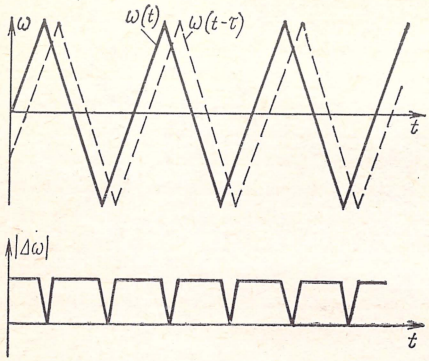


Рис. 2.3. К пояснению действия ЧМ альтиметра

представления о частотной модуляции давно уже разоблачены и отброшены. Ошибка настолько очевидна, что прямо-таки бросается в глаза. Она основана на смешении спектральной и временно́й точек зрения и вытекающем из этого смешения легкомысленном обращении с понятием частоты».

«Дело в том, что периодически модулированное по частоте колебание есть квазипериодический процесс. Следовательно, такой сигнал обладает дискретным спектром с интервалами между линиями, равными частоте модуляции. Но спектр отраженного сигнала имеет точно такой же вид, так как сдвиг по времени не влияет на спектр амплитуд. Следовательно, и спектр результирующего сигнала (т. е.

суммы прямого и отраженного сигналов) будет обладать спектром из так же расположенных дискретных линий, и никаких плавно изменяющихся частот, пропорциональных запаздыванию, не возникает и не может возникнуть, какую бы нелинейную обработку этого спектра мы ни предприняли».

Вслед за этим в упомянутой книге излагается «строгая» теория ЧМ альтиметра, в которой понятие мгновенной частоты не используется. Правда, построить такую теорию исходя из спектрального представления очень трудно. Поэтому используется временно́е представление сигнала, а для упрощения задачи линейная модуляция частоты заменяется модуляцией по гармоническому закону.

Точку зрения А. А. Харкевича вполне можно понять. При подмене частот спектра сигнала его мгновенной частотой возникает много ошибок. Неудивительно, что у ряда исследователей появились предубеждение против использования мгновенной частоты и стремление описывать сигналы только с точки зрения их спектральных составляющих. Посмотрим, к чему ведет такая боязнь нестрогости в использовании спектральных представлений.

2.5. Попробуем быть последовательными

Примем на время точку зрения, выраженную А. А. Харкевичем и еще более определенно Шекелом в приведенных цитатах. При этом нельзя пользоваться представлениями об амплитуде или частоте, которые изменяются во времени. Каждая функция может быть либо строго гармонической, либо негармонической. В последнем случае ее можно описывать некоторым аналитическим выражением, в частности, рядом или интегралом Фурье. Под запрет должно попасть и понятие мгновенного спек-

тра, так как все спектральные составляющие существуют вечно.

Тогда мы должны объявить неграмотными или не имеющими смысла следующие фразы:

«При частотной манипуляции частота излучаемого сигнала меняется скачком в моменты времени, кратные T_0».

«В нашем передатчике используется высокостабильный генератор, частота которого в течение суток меняется не более чем на 3 Гц».

«Спектр сигнала, модулирующего радиовещательный передатчик, изменяется при переходе от речевой передачи к музыкальной».

«Утром я включил генератор гармонического сигнала».

Вместо первой фразы нужно было бы говорить, что «при частотной манипуляции сигнал описывается формулой $s(t)=A_0[\lambda \cos \omega_1 t+(1-\lambda) \cos \omega_2 t]$, где λ — случайная величина, принимающая значение 0 или 1, определяемое передаваемым сообщением и могущее изменяться в моменты времени, кратные T_0».

Вместо второй фразы: «Наш передатчик излучает спектр, часть которого, лежащая вблизи несущей частоты ω_0, сосредоточена в очень узкой полосе» (для количественной характеристики степени стабильности пришлось бы вводить много вспомогательных величин).

Что касается третьей фразы, то мысль о различии спектров речевой и музыкальной передач, кажется, вовсе невозможно выразить, отказавшись от понятия мгновенного спектра.

Последняя фраза оказывается недопустимой потому, что сигнал, начавшийся в некоторый момент времени, не может быть гармоническим.

Еще больше затруднений возникнет при попытке пояснить частотную модуляцию, не прибегая к понятию о мгновенной частоте.

Таким образом, отказ от представления о переменных амплитуде и частоте, хотя и предохраняет от опасности впасть в некоторые заблуждения, приводит к огромным неудобствам при описании простых и привычных явлений. Представление сложных (негармонических) сигналов в форме (2.5) и введение огибающей, мгновенных фазы и частоты чрезвычайно полезны. Необходимо только помнить, что с мгновенной частотой нельзя обращаться как с частотой составляющей спектра, в частности, при изучении прохождения сигнала через линейную цепь.

Примерно так же обстоит дело с понятием о мгновенном спектре, точнее о спектре, изменяющемся во времени. Представление о таком переменном спектре очень наглядно и широко используется инженерами, обычно не задумывающимися о его физическом смысле и математическом выражении. Об этом задумался С. М. Рытов [36], который показал, что сущность такого представления заключается в недоведенном до конца преобразовании Фурье. Грубо говоря, вместо того чтобы рассматривать спектр всего сигнала, заданного на бесконечной оси времени, мы ограничиваемся его спектральным разложением на конечном интервале длительностью T. Ясно, что на различных таких интервалах и «спектры» будут в общем случае различны. Такое представление весьма удобно и адекватно решаемым техническим задачам, если используемые приборы не могут разделять или различать частоты, отличающиеся на величину порядка $1/T$.

2.6. К теории ЧМ альтиметра

Примитивная теория ЧМ альтиметра, подвергнутая критике А. А. Харкевичем, конечно, не может быть признана строгой, но ее достоинством является наглядность.

Можно, и не отказываясь от понятия мгновенной частоты, и используя представление (2.5), а также спектральные представления, изложить теорию ЧМ альтиметра достаточно строго и намного проще, чем это сделано в упомянутой книге [49].

Не останавливаясь на подробностях, покажем, что при сложении двух частотномодулированных сигналов с постоянными амплитудами возникают «биения», т. е. изменения огибающей с частотой, равной разности мгновенных частот. Рассмотрим сумму двух таких сигналов

$$s_1(t) + s_2(t) = A_1 \cos \Phi_1(t) + A_2 \cos \Phi_2(t) \qquad (2.9)$$

с мгновенными частотами $\omega_1 = d\Phi_1/dt$ и $\omega_2 = d\Phi_2/dt$. Пусть для определенности $A_1 \gg A_2$. Обозначив $\Delta(t) = \Phi_2(t) - \Phi_1(t)$, можно записать

$$s_1(t) + s_2(t) = A_1 \cos \Phi_1(t) + A_2 \cos [\Phi_1(t) + \Delta(t)] = A_1 \times$$

$$\times \cos \Phi_1(t) + A_2 \cos \Delta(t) \cos \Phi_1(t) - A_2 \sin \Delta(t) \sin \Phi_1(t) =$$

$$= [A_1 + A_2 \cos \Delta(t)] \cos \Phi_1(t) - A_2 \sin \Delta(t) \sin \Phi_1(t) =$$

$$= \sqrt{[A_1 + A_2 \cos \Delta(t)]^2 + A_2^2 \sin^2 \Delta(t)} \cos [\Phi_1(t) + \Psi(t)],$$

где

$$\Psi(t) = \operatorname{arctg} \frac{-A_2 \sin \Delta(t)}{A_1 + A_2 \cos \Delta(t)}.$$

Огибающая $\sqrt{[A_1 + A_2 \cos \Delta(t)]^2 + A^2_2 \sin^2 \Delta(t)} = \sqrt{A^2_1 + A^2_2 + 2A_1 A_2 \cos \Delta(t)} \approx A_1 [1 + A_2/A_1 \cos \Delta(t)]$ суммарного сигнала изменяется с мгновенной частотой $d\Delta/dt = \omega_2(t) - \omega_1(t)$. Так как эта разность частот значительно меньше значений, принимаемых мгновенной частотой сигнала, огибающую можно легко выделить с помощью детектора и измерить ее частоту. Таким образом, обычное описание ЧМ альтиметра при модуляции частоты в достаточно больших пределах вполне обосновано.

2.7. Существуют ли реально спектральные составляющие?

Вернемся опять к 1930 г., когда Флеминг отрицал реальность существования боковых частот в АМ сигнале. По этому поводу во многих книгах написано, что ошибочное мнение Флеминга было опровергнуто экспериментально, когда боковые частоты были выделены из АМ сигнала резонансными системами. Один из наиболее эффектных демонстрационных опытов был предложен Л. И. Мандельштамом и описан в книге [13]. На обычный язычковый частотомер, используемый для измерения частоты промышленного переменного тока, подадим напряжение от сети, включив последовательно с источником телеграфный ключ. Когда ключ замкнут, резонирует один из язычков, настроенный на частоту 50 Гц, если частота сети соответствует номиналу. Будем теперь периодически прерывать ток ключом, замыкая и размыкая его, например, 2 раза в секунду. Спустя некоторое время наступит установившийся режим, при котором частотомер отметит, помимо несущей частоты 50 Гц, наличие двух боковых частот: 52 Гц и 48 Гц.

Говорят, что этим опытом Флеминг был повержен и реальность существования боковых частот была доказана. Но позвольте, как следует понимать слова «реальность существования»? То, что было доказано описанным опытом и многими другими экспериментами, это возможность выделить из сложного модулированного сигнала простую гармоническую составляющую боковой частоты. Допустим, что нам бы не удалось этого сделать. Тогда боковые частоты следовало бы объявить реально несуществующими? Стать на такую точку зрения

довольно опасно. Покажем это на следующем примере.

Рассмотрим импульс любой формы, длящийся от $t=0$ до $t=T$. Известно, что его спектр представлен интегралом Фурье. На некоторой частоте ω_1 он содержит гармоническую составляющую, определенную на всей оси времени $-\infty<t<\infty$. Мы убеждены, что эта составляющая «реально существует». Значит ли, что ее можно (хотя бы в принципе) выделить из импульса? Если бы это было так, то нарушился бы закон причинности — гармоническая составляющая появилась бы до того, как возник сам импульс[1]).

Таким образом, нужно либо признать, что спектральные составляющие импульса реально не существуют, либо отказаться от отождествления реальности существования с возможностью выделения резонансной системой. По-видимому, целесообразнее считать, что всякое математическое представление сигнала в виде суммы каких-то составляющих, если оно записано верно, определяет реально существующие слагаемые. Эта реальность выражается в возможности пользоваться данным представлением при любых расчетах. Важно уметь выбирать представление, удобное для решаемой задачи.

По этому поводу Л. И. Мандельштам говорил, что число 10 можно с равным основанием представить как 5+5 или 2+8.. Однако для вязальщицы

---

[1]) Если читателя смущает то обстоятельство, что выделяемая гармоническая составляющая имеет бесконечно малую амплитуду, то можно поставить вопрос о выделении части спектра импульса, лежащей в некотором интервале частот (ω_1, ω_2) и имеющей конечную энергию. В общем случае это тоже связано с нарушением закона причинности.

перчаток адекватным является представление $10=5+5$.

Итак, нет никаких сомнений в том, что Флеминг и другие авторы, отрицавшие реальность боковых частот АМ сигнала, заблуждались. Однако нельзя безоговорочно согласиться и с теми из их оппонентов, которые считали, что только, выделив спектральную составляющую резонансной системой, можно доказать ее реальность. Быть может, в аргументациях Флеминга было больше здравого смысла, чем обычно считают.

2.8. Ошибки живучи

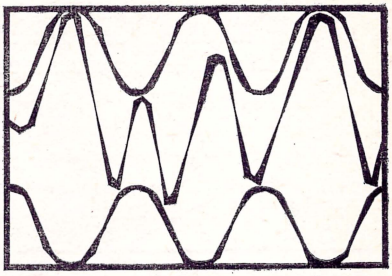
Когда рукопись этой книги уже была подготовлена к печати, автору попался на глаза американский радиолюбительский журнал со статьей [2], в которой существование боковых полос в АМ сигнале хотя и не отвергается, но трактуется как недостаток, вызванный неудачной схемой модулятора. Автор статьи отмечает известный факт, что амплитуда составляющей несущей частоты при АМ остается постоянной, и утверждает, что в обычных модуляторах несущая фактически не модулируется, а используется как опорное колебание для формирования боковых полос.

В статье предлагается «амплитудная модуляция с подавленными боковыми полосами» (?!). Она противопоставляется обычной амплитудной и однополосной модуляции. Зачем, пишет автор статьи, избавляться от несущей, которая не занимает места на шкале частот, и сохранять боковую полосу, имеющую протяженный спектр? Не лучше ли избавиться от боковых и сохранить только несущую, а для передачи информации менять ее амплитуду? Заметьте, это писалось не во времена Флеминга, а в 1974 году!

Далее предлагается схема, которая в первом приближении не отличается от давно известного амплитудного модулятора с регулируемым уровнем несущей. В нем поддерживается практически постоянный коэффициент модуляции, близкий к единице, так как при изменении уровня первичного сигнала пропорционально ему меняется амплитуда несущей. Это дает определенный энергетический выигрыш по сравнению с обыч-

ной АМ, так как в паузах несущая не передается и ее средний уровень значительно снижается. Достигается этот выигрыш ценой существенных искажений. Что же касается спектра полученного сигнала, то он, конечно, содержит, помимо несущей, две боковые полосы. К обычным боковым частотам здесь добавлены еще более близкие к несущей боковые, образованные инфразвуковой модуляцией несущей по закону изменения огибающей первичного сигнала. Автор же статьи без каких-либо доказательств утверждает, что его сигнал вовсе не содержит боковых частот, а состоит из одной несущей с переменной амплитудой и поэтому имеет нулевую ширину спектра. Здесь, пишет он, «полоса частот является только функцией селективности приемника».

Итак, автор статьи повторяет буквально то, что писал Флеминг и что было опровергнуто 50 лет назад.



3

О КОМПЛЕКСНОМ ПРЕДСТАВЛЕНИИ СИГНАЛА

3.1. Обобщение символического метода

Гармонический сигнал $u(t)=A_0\cos(\omega_0 t+\varphi_0)$ можно записать в экспоненциальной форме

$$u(t)=\frac{A_0}{2}\left(e^{i(\omega_0 t+\varphi_0)}+e^{-i(\omega_0 t+\varphi_0)}\right), \qquad (3.1)$$

либо

$$u(t)=\mathrm{Re}\left\{A_0 e^{i(\omega_0 t+\varphi_0)}\right\}. \qquad (3.2)$$

При механической интерпретации выражение (3.1) означает замену колебательного движения вдоль прямой двумя вращательными движениями с постоянными угловыми скоростями $+\omega_0$ и $-\omega_0$ (рис. 3.1), а выражение (3.2) — представление колебательного движения в виде проекции одного вращательного движения с положительной угловой скоростью ω_0 (рис. 3.2). Выражение в фигурных скобках (3.2) является комплексной функцией действительной переменной

$$\dot{u}(t)=A_0 e^{i(\omega_0 t+\varphi_0)}=A_0[\cos(\omega_0 t+\varphi_0)+i\sin(\omega_0 t+\varphi_0)], \qquad (3.3)$$

действительная часть которой совпадает с исходным гармоническим сигналом $u(t)$, а мнимая часть $v(t)=A_0 \sin(\omega_0 t+\varphi_0)$ отличается от исходного сиг-

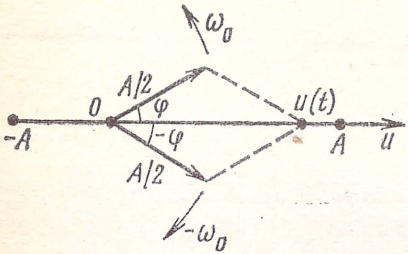


Рис. 3.1. Колебание как сумма двух вращений

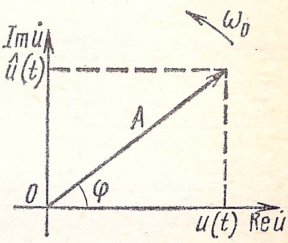


Рис. 3.2. Колебание как проекция вращения

нала поворотом фазы на $-\pi/2$ и может быть названа *сопряженным гармоническим сигналом.*

При решении многих задач теории цепей удобно вместо гармонического сигнала рассматривать комплексный сигнал $\dot{u}(t)$ или его комплексную амплитуду $\dot{A}=A_0 e^{i\varphi_0}$. Как известно, метод решения задач путем замены действительной гармонической функции комплексной амплитудой называется *символическим методом* и находит широкое применение.

Обобщением символического метода является представление сложного (т. е. не гармонического) сигнала $u(t)$ в виде действительной части комплексного сигнала

$$\dot{u}(t)=u(t)+iv(t), \qquad (3.4)$$

где $v(t)$ — пока не определенный сигнал, сопряженный с $u(t)$. Переписав (3.4) в экспоненциальной форме

$$u(t) = A(t)\,e^{i\Phi(t)} = A(t)\,[\cos\Phi(t) + i\sin\Phi(t)], \qquad (3.5)$$

где, очевидно,

$$A(t)\cos\Phi(t) = u(t),\ A(t)\sin\Phi(t) = v(t), \qquad (3.6)$$

можно определить *огибающую* сигнала,

$$A(t) = [u^2(t) + v^2(t)]^{1/2}, \qquad (3.7)$$

его *полную мгновенную фазу*

$$\Phi(t) = \operatorname{arctg}(v(t)/u(t)) \qquad (38)$$

и *мгновенную частоту*

$$\omega(t) = \frac{d\Phi}{dt} = \frac{v'(t)\,u(t) - u'(t)\,v(t)}{u^2(t) + v^2(t)}, \qquad (3.9)$$

где штрихи обозначают производные по t.

Наконец, задавшись некоторой средней частотой ω_0 и определив *мгновенную начальную фазу*

$$\varphi(t) = \Phi(t) - \omega_0 t, \qquad (3.10)$$

можно записать исходный сложный сигнал $u(t)$ в квазигармонической форме, аналогичной (3.1):

$$u(t) = A(t)\cos[\omega_0 t + \varphi(t)] \qquad (3.11)$$

или в форме

$$u(t) = \operatorname{Re}\{\dot{A}(t)\,e^{i\omega_0 t}\}, \qquad (3.12)$$

где $\dot{A}(t) = A(t)\,e^{i\varphi(t)}$ — *комплексная огибающая* сигнала.

Замена сигнала $u(t)$ его комплексной огибающей удобна при решении многих задач и представляет собой распространение символического метода на сложные сигналы.

3.2. Однозначно ли определены огибающая и мгновенная частота?

Прежде всего заметим, что мгновенная начальная фаза $\varphi(t)$ и комплексная огибающая $\dot{A}(t)$ зависят

от выбора средней частоты ω_0, а поскольку ω_0 задается произвольно, то об однозначности определения $\varphi(t)$ и $A(t)$ говорить не приходится. Однако действительная огибающая $A(t)$, мгновенная полная фаза $\Phi(t)$ и мгновенная частота $\omega(t)$ от выбора ω_0 не зависят и целиком определяются функциями $u(t)$ и $v(t)$ по ф-лам (3.7), (3.8) и (3.9). В частности, $\omega(t)=\omega_0+d\varphi/dt$ не зависит от ω_0, так как при изменении ω_0 соответственно изменяется и $d\varphi/dt$. Поэтому если установить некоторое правило, по которому из сигнала $u(t)$ определяется сопряженный сигнал $v(t)$, то тем самым будут однозначно определены $A(t)$, $\Phi(t)$ и $\omega(t)$. Заметим, что если из четырех функций — $u(t)$, $v(t)$, $\Phi(t)$ и $A(t)$ — задать две любые, то остальные определяются однозначно, по ур-ниям (3.6).

Таким образом, вопрос об однозначном определении огибающей, мгновенной полной фазы и мгновенной частоты сигнала сводится к определению сопряженного сигнала $v(t)$ по сигналу $u(t)$. Вероятно читателю известно, что обычно сопряженный сигнал выбирают как преобразование Гильберта от $u(t)$. Впрочем, если читатель этого не знает, то он может спокойно продолжать чтение, так как все необходимые сведения о преобразовании Гильберта будут далее изложены. Почти во всех работах, где этот вопрос рассматривается, отмечают, что огибающая, мгновенные фаза и частота удовлетворяют некоторым естественным и удобным требованиям только при том условии, что функция $v(t)$ сопряжена с $u(t)$ по Гильберту. Это, вообще говоря, верно. Но при перечислении указанных требований многие авторы совершают грубую ошибку, заслуживающую упоминания на этих страницах. Сейчас уже трудно сказать, кто ошибся первым. Но в течение более десяти лет в ряде учебников и моно-

графий переписывается с небольшими вариациями следующее утверждение.

Для заданной функции $u(t)$ огибающую $A(t)$ в в (3.1) можно определить различными способами. Потребуем, чтобы функция $A(t)$ удовлетворяла следующим двум условиям, которые, собственно говоря, и оправдывают термин «огибающая»:

1) при любом значении t, $A(t) \geqslant |u(t)|$; 2) при тех значениях t, при которых предыдущее неравенство обращается в равенство, т. е. огибающая $A(t)$ и функция $|u(t)|$ имеют общую точку, их производные также совпадают, т. е. их графики имеют общую касательную.

Тогда огибающая, удовлетворяющая этим условиям, определяется однозначно, как модуль комплексного сигнала $\dot{u}(t)$ (3.4), где $v(t)$ — преобразование Гильберта от $u(t)$.

Приведенное утверждение глубоко ошибочно. Перечисленные два условия удовлетворяются не только при выборе в качестве мнимой части комплексного сигнала преобразования Гильберта от $u(t)$, но и при совершенно произвольном выборе любой непрерывной функции в качестве $v(t)$. В этом легко убедиться. По определению огибающей

$$A(t) = |\dot{u}(t)| = \sqrt{u^2(t) + v^2(t)} = |u(t)| \times$$
$$\times \sqrt{1 + v^2(t)/u^2(t)} \geqslant |u(t)|. \qquad (3.13)$$

Здесь везде радикалы обозначают арифметическое значение корня. Неравенство (3.13) обращается в равенство при тех значениях t, для которых $v(t) = 0$. Производная огибающей

$$A'(t) = \frac{d}{dt}\sqrt{u^2(t) + v^2(t)} = \frac{u(t)\,u'(t) + v(t)\,v'(t)}{\sqrt{u^2(t) + v^2(t)}}. \qquad (3.14)$$

Положив в (3.14) $v(t)=0$, найдем, что в точках совпадения $A(t)$ и $|u(t)|$

$$A'(t) = |u(t)|'. \tag{3.15}$$

Таким образом, для выполнения условий 1) и 2) вовсе не требуется, чтобы непрерывные функции $u(t)$ и $v(t)$ были связаны преобразованием Гильберта.

Почему же, в таком случае, в качестве мнимой части комплексного сигнала $\dot{u}(t)$ почти все авторы используют преобразование Гильберта от $u(t)$? Этому вопросу и посвящена в основном настоящая глава[1].

Короткий ответ на этот вопрос заключается в том, что использование преобразования Гильберта вместо произвольной функции $v(t)$ позволяет, помимо условий 1 и 2, выполнить и ряд других условий, при которых удобнее пользоваться огибающей и мгновенной частотой. Впрочем, далеко не все пожелания, касающиеся естественности и удобства огибающей и мгновенной частоты, удается выполнить с помощью преобразования Гильберта. Попробуем разобраться в этом подробнее.

Предварительно введем одно определение. Будем называть комплексный сигнал (3.4) гильбертовским, если в качестве сопряженного сигнала $v(t)$ выбрана функция $\hat{u}(t)$, являющаяся прямым преобразованием Гильберта от $u(t)$.

3.3. Что такое преобразование Гильберта?

Здесь пора выполнить обещание и напомнить читателю, что такое преобразование Гильберта. Пусть сигнал $u(t)$ интегрируем в квадрате в бесконечных пределах, т. е. несобственный

[1] Многие мысли, изложенные в данной главе, заимствованы из статьи [6].

интеграл $\int_{-\infty}^{\infty} u^2(t)dt$ сходится. Тогда, как известно, существует преобразование Фурье сигнала $u(t)$, равное

$$S(\mathrm{i}\omega) = \int_{-\infty}^{\infty} u(t)\mathrm{e}^{-\mathrm{i}\omega t}\,dt, \tag{3.16}$$

называемое спектральной плотностью сигнала и позволяющее представить $u(t)$ в спектральной форме

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\mathrm{i}\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,d\omega, \tag{3.17}$$

Спектральная плотность $S(\mathrm{i}\omega)$ действительного сигнала $u(t)$ является эрмитовской, т. е. ее действительная часть является четной, а мнимая часть — нечетной функцией от ω. Это видно из (3.16), так как при изменении знака ω действительная часть интеграла остается неизменной, а мнимая часть меняет знак.

Рассмотрим функцию

$$\widetilde{S}(\mathrm{i}\omega) = \begin{cases} -\mathrm{i}\,S(\mathrm{i}\omega) & \omega > 0, \\ 0 & \omega = 0, \\ \mathrm{i}\,S(\mathrm{i}\omega) & \omega < 0. \end{cases} \tag{3.18}$$

которая, очевидно, также является эрмитовской. Тогда ее обратное преобразование Фурье

$$\hat{u}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{S}(\mathrm{i}\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,d\omega \tag{3.19}$$

и называется преобразованием Гильберта от $u(t)$.

Другими словами, $u(t)$ и $\hat{u}(t)$ имеют одинаковые модули спектральных плотностей, а аргумент (фаза) спектральной плотности $\widetilde{S}(\mathrm{i}\omega)$ сдвинута относительно $S(\mathrm{i}\omega)$ на $-\pi/2$ в области положительных частот и на $\pi/2$ в области отрицательных частот. Таким образом, преобразование Гильберта является естественным обобщением мнимой части в ф-ле (3.3).

Можно и прямо определить прямое и обратное интегральное преобразование Гильберта, минуя преобразование

Фурье, а именно:

$$\left.\begin{aligned} \hat{u}(t) &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{u(\tau)}{t-\tau}\,d\tau, \\ u(t) &= -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\hat{u}(\tau)}{t-\tau}\,d\tau, \end{aligned}\right\} \qquad (3.20)$$

где интегралы понимаются в смысле главного значения Коши. Формулы (3.21) легко выводятся из (3.19) и (3.20). Для дальнейшего важно понимать, что «сопряженные по Гильберту» сигналы $u(t)$ и $\hat{u}(t)$ имеют спектральные плотности, отличающиеся только поворотом фазы на $\pm\pi/2$.

Заметим, что вместо традиционной записи преобразований Гильберта (3.21) удобнее пользоваться другими формулами, получающимися при замене переменных $t-|\tau|=x$:

$$\left.\begin{aligned} \hat{u}(t) &= \frac{1}{\pi}\lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty}\frac{u(t-x)-u(t+x)}{x}\,dx, \\ u(t) &= \frac{1}{\pi}\lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty}\frac{\hat{u}(t+x)-\hat{u}(t-x)}{x}\,dx, \end{aligned}\right\} \qquad (3.21)$$

Определение преобразования Гильберта (3.21) или (3.20) можно теперь распространить на более широкий класс сигналов, чем интегрируемые в квадрате, в частности, на периодические и почти периодические сигналы[1]. Например, для гармонического сигнала $u(t)=A\cos(\omega_0 t+\varphi)$ сопряженный по Гильберту сигнал легко найти из (3.21):

$$\hat{u}(t) = \frac{A}{\pi}\lim_{\varepsilon\to 0}\int_{\varepsilon}^{\infty}\frac{\cos[\omega_0(t-x)+\varphi]-\cos[\omega_0(t+x)+\varphi]}{x}\,dx=$$

[1] Некоторые достаточные (но не необходимые) условия существования ограниченного преобразования Гильберта сформулированы в работе [45].

$$= \frac{2A}{\pi} \sin(\omega_0 t + \varphi) \lim_{\varepsilon \to 0} \int_{\varepsilon}^{\infty} \frac{\sin \omega_0 x}{x} dx = A \sin(\omega_0 t + \varphi), \tag{3.22}$$

так как $\int_0^{\infty} \frac{\sin \omega_0 x}{x} dx = \frac{\pi}{2}$.

Заметим, что при этом выводе молчаливо предполагалось, что $\omega_0 > 0$. В противном случае $\int_0^{\infty} \frac{\sin \omega_0 x}{x} dx = -\frac{\pi}{2}$ и

$\hat{u}(t) = -A \sin(\omega_0 t + \varphi) = A \sin(|\omega_0| t - \varphi)$.

Итак, для гармонического сигнала преобразование Гильберта сводится к повороту начальной фазы на $-\pi/2$. Отсюда следует, что огибающая гармонического сигнала $\sqrt{u^2(t) + \hat{u}^2(t)} = A$ совпадает с его обычной амплитудой, а мгновенная частота $\frac{d}{dt} \operatorname{arctg} \frac{\hat{u}(t)}{u(t)} = \omega_0$ совпадает с частотой его единственной спектральной составляющей.

Таким образом, гильбертовское определение огибающей удовлетворяет естественному условию:

3) огибающая должна быть определена так, чтобы для гармонического сигнала она совпадала с обычной его амплитудой, а мгновенная частота — с обычной частотой[1].

Вот это условие, как легко проверить, выполняется не при любом выборе сопряженной функции $v(t)$. Теперь уже видно явное преимущество гильбертовского сигнала, у которого $v(t) = \hat{u}(t)$, перед другими комплексными представлениями с произвольным выбором сопряженного сигнала $v(t)$. Однако является ли гильбертовское преобразование единственным, удовлетворяющим условию 3?

[1] Заметим, между прочим, что если определить сопряженный сигнал не как прямое, а как обратное преобразование Гильберта от $u(t)$, то огибающая для гармонического сигнала по-прежнему равнялась бы A, но мгновенная частота была бы отрицательной и равной $-\omega_0$. В ф-ле (3.18) знаки выбраны так, чтобы для гильбертовского гармонического сигнала $\omega(t) = \omega_0$.

3.4. Другие определения сопряженного сигнала

На поставленный выше вопрос приходится ответить отрицательно. Можно различными способами определить сопряженный сигнал $v(t)$, чтобы вытекающие из этого определения значения огибающей и мгновенной частоты удовлетворяли условию 3. Многие авторы прямо или косвенно предлагали для узкополосных сигналов $u(t)$ следующее комплексное представление:

$$\dot{u}(t) = u(t) + i\,u'(t)\sqrt{-u(t)/u''(t)}. \qquad (3.23)$$

При таком определении сопряженного сигнала огибающая

$$A(t) = \sqrt{u^2 + v^2} = \sqrt{u^2 - (u')^2\,u/u''} \qquad (3.24)$$

(для упрощения записи аргумент t опущен).

Легко проверить, что для любого гармонического сигнала отсюда получаются в полном соответствии с условием 3 значения $A(t)=A$ и $\omega(t)=\omega_0$.

Несмотря на это, преобразование (3.23) не удовлетворяет другим естественным требованиям. Дело в том, что при незначительных изменениях $u(t)$ функция $v(t)$ может изменяться в огромных пределах. Например, заменим в гармоническом сигнале $u(t)$ на небольшой доле периода синусоиду прямой линией, как показано пунктиром на рис. 3.3. Это может быть, например, вызвано добавлением очень слабой помехи и почти не отражается на ходе $u(t)$. В то же время на этом участке $u''(t)=0$, так что $v(t)$, а с ней и «огибающая» принимают бесконечное значение. Более того, при ничтожно малых приращениях $u(t)$ вторая производная может изменить знак и тогда в (3.25) вообще нельзя извлечь корня.

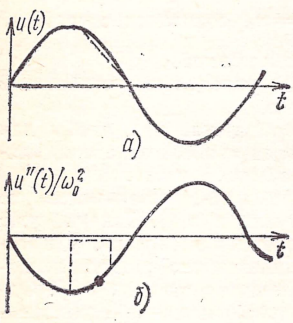


Рис. 3.3. Слегка измененная синусоида *(а)* и ее вторая производная *(б)*

В соответствии с этими сопряжениями в [6] введены два условия, которым должны удовлетворять огибающая, мгновенная фаза и мгновенная частота для того, чтобы этими понятиями можно было пользоваться на практике:

4) малым (в смысле среднего квадратичного отклонения) изменениям исходного сигнала $u(t)$ должны соответствовать также малые изменения $A(t)$ и $\omega(t)$;

5) мгновенные фаза и частота не должны зависеть от мощности сигнала.

В [6] показано, что из условий 4 и 5 следует линейность оператора преобразования $u(t)$ в $v(t)$. Единственным же линейным оператором, при котором для всех гармонических сигналов выполняется условие 3, является преобразование Гильберта.

В. И. Тихонов [40] предложил определять сопряженный сигнал следующим образом:

$$v(t) = -(1/\omega_0)(du/dt), \qquad (3.25)$$

где ω_0 — средняя частота сигнала.

Это преобразование также является линейным и удовлетворяет условиям 4 и 5. Что же касается условия 3, то оно выполняется только для гармонического сигнала с заданной частотой ω_0. При переходе к гармоническому сигналу с другой частотой нужно в (3.25) изменить коэффициент, а для этого частоту нужно знать заранее. Если распространять это определение на негармонические сигналы, то возникают затруднения в выборе ω_0. Тем не менее часто утверждают, что для достаточно узкополосных сигналов средняя частота выражена ярко и никаких затруднений в применении ф-лы (3.27) не возникает.

Многие авторы, например [11, 24, 30] и др., полагают, что для широкополосных сигналов вообще не следует говорить об огибающей и мгновенной частоте. Эта точка зрения подкрепляется тем обстоятельством, что для узкополосных сигналов любой инженер может практически однозначно построить огибающую графически на основе ее свойств 1 и 2, как показано на рис. 3.4, тогда как для широкополосного сигнала (например, рис. 3.5) такое построение весьма затруднительно.

Тем не менее существуют две причины, не позволяющие отказаться от представления широкополосных сигналов в форме (3.11). *Во-первых,* нельзя установить разумную границу между узкополосными сигналами, у которых огибающая существует, и широкополосными, не имеющими ее. *Во-вторых,* отказ от комплексного представления широкополосных сигналов заставит отказаться от простой и изящной теории однополосной модуляции, а также затруднит понимание процесса амплитудной и частотной модуляции, детектирования, преобразования частоты и т. д.

Заметим еще, что оба определения — (3.21) и (3.25) — в применении к узкополосным сигналам дают почти одинаковые значения сопряженного сигнала, причем различие между ними тем меньше, чем уже спектр сигнала. Здесь имеется в виду, конечно, относительная ширина спектра, т. е. отношение

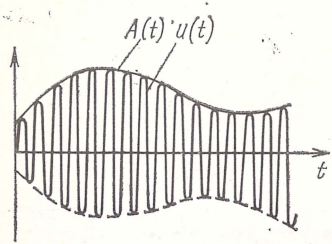


Рис. 3.4. Узкополосный сигнал и его огибающая

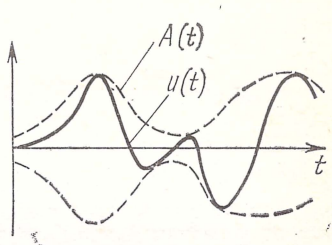


Рис. 3.5. Построить огибающую широкополосного сигнала трудно

ее абсолютной ширины[1]) к средней частоте. В предельном случае при ширине спектра, равной нулю, т. е. для гармонического сигнала, оба определения, как уже отмечалось, совпадают.

Предлагались и некоторые другие определения огибающей и мгновенной частоты. Примером может служить упоминавшееся в предыдущей главе определение мгновенной частоты (2.8), предложенное Шекелом. Как уже отмечалось, из определения мгновенной частоты непосредственно вытекает определение «сопряженного» сигнала и огибающей. Как ни странно, это определение, недостатки которого отмечал и предложивший его Шекел, время от времени предлагается вновь как алгоритм для измерения мгновенной частоты негармонического сигнала.

Еще одно определение огибающей для сигнала со средней частотой спектра ω_0 основано на введении в качестве

---

[1]) Под абсолютной шириной спектра можно понимать, например, ширину области частот, в которой сосредоточено 95% энергии или мощности сигнала. Впрочем, возможны и любые другие разумные определения ширины спектра.

«сопряженного» сигнала $v(t)$, отличающегося от $u(t)$ только сдвигом во времени:

$$v(t) = u(t - t_c), \qquad (3.26)$$

где $t_c = \pi/2\omega_0$ — четверть периода средней частоты. При этом определении огибающая $A(t) = \sqrt{u^2(t) + u^2(t - t_c)}$.

3.5. Конкурс определений

Какое же из определений сопряженного сигнала является наиболее удобным? Все рассматриваемые определения для узкополосных сигналов согласуются с интуитивными представлениями об огибающей и о мгновенной частоте. Сосредоточим внимание на индивидуальных особенностях, присущих каждому из определений.

Прежде всего отметим, что определения (3.23) и (3.25) обладают очень удобным свойством — *локальностью*. При этих определениях сопряженный сигнал, а следовательно, и огибающая и мгновенная частота в некоторой точке t, в которой функция $u(t)$ непрерывна и дифференцируема, однозначно определяются значениями функции $u(t)$ в сколь угодно малом интервале около этой точки. В частности, если на конечном интервале времени $u(t) = 0$, то по определению В. И. Тихонова и $v(t) = 0$ на том же интервале.

Что же касается гильбертовского определения сопряженного сигнала, то оно не является локальным. Как видно из (3.21) или (3.22), если сопряженный сигнал $v(t)$ по определению положить равным $\hat{u}(t)$, то для вычисления его значения в одной точке t необходимо, в принципе, знать всю функцию $u(t)$, $-\infty < t < \infty$. Правда, из (3.22) видно, что модуль подынтегральной функции велик лишь при значениях x, близких к нулю, и поэтому существенную роль при определении $\hat{u}(t)$ в данной точке играет поведение функции $u(t)$ только вбли-

зи этой точки. Такая «относительная» локальность проявляется тем больше, чем узкополоснее сигнал $u(t)$. Однако строгой локальности здесь нет и в частности, $v(t)$ может принимать самые различные значения при $u(t)=0$.

Почему же, несмотря на это, подавляющее большинство исследователей пользуются более сложным и менее наглядным преобразованием Гильберта вместо простой ф-лы (3.25)? Не является ли это просто данью традиции?

Нет, не является. Определение сопряженного сигнала и огибающей на основании преобразования Гильберта обладает важными преимуществами, ради которых стоит пожертвовать наглядностью представления широкополосных сигналов. Помимо отмеченных выше пяти свойств параметров комплексного представления сигнала, укажем еще два, которым удовлетворяют только гильбертовские сигналы:

6) для того чтобы комплексный сигнал был гильбертовским, необходимо и достаточно, чтобы его спектральная плотность тождественно равнялась нулю при $\omega<0$;

7) при умножении гильбертовского сигнала на $e^{i\psi}$, где ψ — произвольный угол, или на $e^{i\nu t}$, где $\nu>0$, результирующий сигнал остается гильбертовским.

Предоставим читателю самостоятельно убедиться, что свойство 7 вытекает из 6. Несколько сложнее показать, что из 7 следует 6. Это сделано, по существу, в [21], где поставлена задача отыскания общей огибающей семейства кривых, представляющих результат сдвига начальных фаз всех спектральных составляющих сигнала на любой угол ψ, и показано, что эта огибающая равна модулю гильбертовского сигнала. Таким образом, требование

инвариантности огибающей (или мгновенной частоты) при любых изменениях начальной фазы сигнала оказывается достаточным для того, чтобы гильбертовский сигнал был единственным возможным комплексным представлением сигнала.

Из свойств 6 и 7 вытекает, что всякий сдвиг начальных фаз сигнала можно выразить умножением гильбертовского сигнала на $e^{i\varphi}$, а всякое преобразование частоты (транспонирование спектра) вверх, в частности однополосную модуляцию, можно выразить умножением гильбертовского сигнала на $e^{i\nu t}$. Именно это позволяет представить сигнал $s(t)$, являющийся результатом однополосной модуляции первичного (низкочастотного) сигнала $x(t)$, аналитически, минуя преобразование Фурье:

$$s(t) = \mathrm{Re}\left\{ \dot{x}(t)\, e^{i\nu t} \right\} = \mathrm{Re}\left\{ x(t)\, e^{i\nu t} + i\hat{x}(t)\, e^{i\nu t} \right\} =$$
$$= x(t)\cos\nu t - \hat{x}(t)\sin\nu t, \qquad (3.27)$$

где ν — несущая частота при однополосной модуляции; $\hat{x}(t)$ — преобразование Гильберта от $x(t)$; $\dot{x}(t) = x(t) + i\hat{x}(t)$ — гильбертовский (комплексный) сигнал.

Ни при каком другом определении сопряженного сигнала ф-ла (3.27) не описывает однополосную модуляцию. Действительно, спектр комплексного сигнала $\dot{x}(t)\, e^{i\nu t}$ будет лежать выше несущей частоты ν тогда и только тогда, когда спектр $\dot{x}(t)$ отличен от нуля лишь при $\omega > 0$, а это согласно свойству 6 характеризует гильбертовский сигнал.

Из ф-лы (3.27) следует, между прочим, известный способ осуществления однополосной модуляции без применения фильтров (фазокомпенсационный метод однополосной модуляции).

Аналогично применение гильбертовских сигналов позволяет построить стройную теорию преобра-

зования частоты, при котором огибающая сохраняется неизменной, а мгновенная частота смещается на заданную величину. При построении теории оптимальной обработки сигналов в каналах с флуктуациями фазы и аддитивным шумом, в частности в различных дисперсивных каналах, только использование гильбертовского представления позволяет получить достаточно простые и наглядные алгоритмы и расчетные формулы. Это также обусловлено свойствами 6 и 7 гильбертовских сигналов.

Перечисленные возможности, возникающие только при использовании гильбертовских сигналов, убеждают в том, что именно гильбертовское представление комплексного сигнала должно применяться для определения огибающей и мгновенной частоты сигнала. Ради этих возможностей можно примириться и с уже известными недостатками гильбертовских сигналов, а именно с отсутствием свойства локальности сопряженного сигнала и несохранением финитности при преобразовании Гильберта.

3.6. Как же быть с финитными сигналами?

Последний недостаток гильбертовского представления сигналов, собственно говоря, даже не является недостатком, а вытекает из природы вещей. Если некоторый строго финитный сигнал подается на вход канала, в котором фазы всех составляющих спектральной плотности поворачиваются на одинаковый угол ψ, не кратный π, то на выходе канала сигнал оказывается нефинитным. Поэтому если к определению огибающей предъявлено требование инвариантности относительно сдвига фаз спектральной плотности, то огибающая любого сигнала не может быть финитной.

Тем не менее при исследовании синхронных систем передачи дискретных сообщений, когда все реализации сигнала финитны и имеют одинаковую длительность T, кажется неестественным рассматривать нефинитные сопряженные сигналы и нефинитные огибающие. Выход из этого положения найден

довольно давно. При рассмотрении синхронных систем сигнал $s(t)$, заданный на отрезке $(0, T)$ и интегрируемый с квадратом на этом отрезке, периодически продолжается в обе стороны. Для получившегося периодического сигнала $u(t)$ преобразованием Гильберта является также периодический сигнал $\hat{u}(t)$ с тем же периодом T. Один период сигнала $\hat{u}(t)$ на интервале $(0, T)$ можно принять за сигнал, сопряженный с финитным сигналом $s(t)$. Если теперь определить огибающую $s(t)$ на основе этого сопряженного сигнала, то она также окажется финитной. Такое «модифицированное» преобразование Гильберта использовано в [46]. Более строгая и изящная теория такого преобразования дана в статье В. И. Коржика [22][1]. Сущность этого обобщенного преобразования, которое можно назвать преобразованием Гильберта—Коржика, заключается в следующем. Сигнал $u(t)$, заданный на отрезке $(0, T)$, можно представить в виде ряда Фурье в комплексной форме

$$(t) = \sum_{k=-\infty}^{\infty} \dot{c}_k \, e^{ik\Omega t}, \tag{3.28}$$

где $\Omega = 2\pi/T$.

Тогда преобразованием Гильберта—Коржика от $u(t)$ является сигнал

$$\hat{u}(t) = \sum_{k=-\infty}^{\infty} \hat{c}_k \, e^{ik\Omega t}, \tag{3.29}$$

где $\hat{c}_k = \begin{cases} -i\dot{c}_k & \text{при } k > 0, \\ 0 & \text{при } k = 0, \\ i\dot{c}_k & \text{при } k < 0. \end{cases}$

Комплексный сигнал, построенный на основе преобразования Гильберта—Коржика, удовлетворяет условиям 1, 2, 4, 5. Что же касается условий 3, 6, 7, то они удовлетворяются в слегка измененном виде, а именно:

3′) огибающая отрезка гармонического сигнала $A\cos(\omega_0 t + \varphi)$, $0 < t < T$, совпадает с его амплитудой A, а мгновенная частота — с частотой ω_0, если ω_0 кратна $2\pi/T$;

---

[1] В [22] преобразование Гильберта распространено и на дискретное время, а также получены более общие результаты.

6′) для того чтобы комплексный сигнал, заданный на отрезке, был гильбертовским (в смысле обобщенного преобразования), необходимо и достаточно, чтобы коэффициенты его ряда Фурье равнялись нулю для всех отрицательных гармоник;

7′) при умножении гильбертовского сигнала, заданного на отрезке $(0, T)$, на $e^{i\psi}$, где ψ — произвольный угол, или на $e^{i\nu t}$, где ν — положительная частота, кратная $2\pi/T$, результирующий сигнал остается гильбертовским на том же отрезке.

Различие между классическим преобразованием Гильберта и преобразованием Гильберта—Коржика состоит, по существу, в том, что первое определено на пространстве сигналов $L_2(-\infty, \infty)$, т. е. для сигналов, заданных на всей оси времени и интегрируемых на этой оси в квадрате, тогда как второе определено на пространстве $L_2(0, T)$, т. е. для сигналов, заданных и интегрируемых в квадрате на отрезке $(0, T)$.

Для того чтобы яснее представить различие между классическим преобразованием Гильберта и преобразованием Гильберта—Коржика, рассмотрим ограниченный сигнал $u(t)$, отличный от нуля только на интервале $(0, T)$[1]. Его можно рассматривать как заданный на бесконечной прямой $(-\infty, \infty)$, но равный нулю при $t<0$ и $t>T$. В этом случае сигнал интегрируем в квадрате на всей оси и к нему применимо классическое преобразование Гильберта. Определим с его помощью сопряженный сигнал $\hat{u}_\infty(t)$ и огибающую $A_\infty(t)$. Обе эти функции, как уже отмечалось, не являются финитными.

Но тот же сигнал $u(t)$ можно рассматривать как заданный на интервале $(0, T)$ и определить сопряженный сигнал $\hat{u}_T(t)$ с помощью преобразования Гильберта—Коржика, а затем и огибающую $A_T(t)$.

Обе различные огибающие $A_\infty(t)$ и $A_T(t)$ являются инвариантными относительно сдвига фаз всех спектральных составляющих сигнала $u(t)$, но в разном смысле. Для $A_\infty(t)$ инвариантность имеет место, если сдвигать фазу на всех частотах спектральной плотности, т. е. заменить $S(i\omega)$ на $S(i\omega)e^{i\psi\,\mathrm{sgn}\,\omega}$. Для $A_T(t)$ инвариантность выполняется, если сдвинуть на угол ψ фазы составляющих ряда Фурье, т. е. в (3.29) умножить все $\dot{c}_k$ на $e^{i\psi\,\mathrm{sgn}\,k}$.

В отличие от «огибающей», определенной по Тихонову, $A_\infty(t)$ и $A_T(t)$ не являются локальными характеристиками сигнала, хотя функция $A_T(t)$ финитна.

---

[1] Этот интервал, разумеется, можно было бы заменить любым другим интервалом времени с длительностью T.

Для относительно узкополосного сигнала $u(t)$ различие между $A_\infty(t)$ и $A_T(t)$ заметно только вблизи границ интервала $(0, T)$. Чем меньше отношение ширины спектра сигнала к средней частоте, тем ближе друг к другу огибающие $A_\infty(t)$ и $A_T(t)$. Поэтому на практике часто не оговаривают, какая из огибающих имеется в виду, а иногда над этим и не задумываются, что также может приводить к досадным ошибкам.

3.7. Занимательная задача

Чаще всего огибающая, определенная на основании преобразования Гильберта, хорошо согласуется с интуитивным представлением инженеров. Рассмотрим сигнал $u(t) = \cos\omega_1 t \cos\omega_2 t$ и найдем его огибающую. Предварительно попробуем определить ее интуитивно.

Предположим, что $\omega_2 \gg \omega_1$. Тогда наш сигнал выглядит так, как показано на рис. 3.6, и огибающая, показанная пунктиром, описывается формулой $A(t) = |\cos\omega_1 t|$. Аналогично должна определяться огибающая и в случае, когда $\omega_1 \gg \omega_2$, но теперь она будет равна $A(t) = |\cos\omega_2 t|$. Если же частоты ω_1 и ω_2 близки друг к другу, то интуиция отказывает. Во всяком случае, многие инженеры, которым я задавал этот вопрос, становились в тупик.

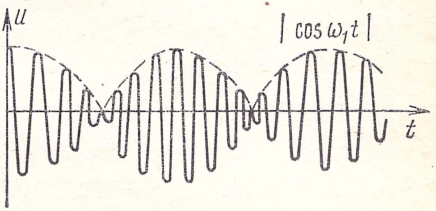


Рис. 3.6. Произведение двух гармонических функций и его огибающая

Решим теперь эту задачу с помощью преобразо-

вания Гильберта. Сигнал $u(t)$ представим в виде

$$u(t) = 0{,}5\cos(\omega_1 + \omega_2)t + 0{,}5\cos(\omega_1 - \omega_2)t. \qquad (3.30)$$

Учитывая (3.22) и замечание, приведенное после этой формулы, сопряженный сигнал запишем в виде

$$u(t) = 0{,}5\sin(\omega_1 + \omega_2)t + 0{,}5\sin|\omega_1 - \omega_2|t. \qquad (3.31)$$

Отсюда квадрат огибающей

$$A^2(t) = u^2(t) + \hat{u}^2(t) = 0{,}25[\cos^2(\omega_1 + \omega_2)t +$$
$$+ \cos^2(\omega_1 - \omega_2)t + \sin^2(\omega_1 + \omega_2)t + \sin^2(\omega_1 - \omega_2)t +$$
$$+ 2\cos(\omega_1 + \omega_2)t\cos(\omega_1 - \omega_2)t \pm 2\sin(\omega_1 + \omega_2)t \times$$
$$\times \sin(\omega_1 - \omega_2)t], \qquad (3.32)$$

где знак «плюс» следует брать при $\omega_1 > \omega_2$, а «минус» — при $\omega_1 < \omega_2$. Таким образом, при $\omega_2 > \omega_1$

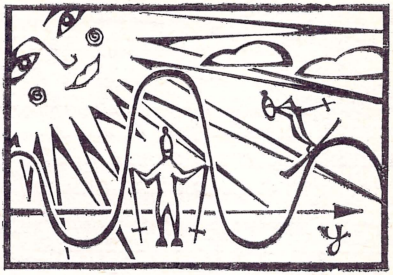
$$A^2(t) = 0{,}5\{1 + \cos[(\omega_1 + \omega_2)t + (\omega_1 - \omega_2)t]\} =$$
$$= 0{,}5(1 + \cos 2\omega_1 t) = \cos^2\omega_1 t \text{ и } A(t) = \sqrt{\cos^2\omega_1 t} =$$
$$= |\cos\omega_1 t|, \qquad (3.33)$$

при $\omega_1 > \omega_2$

$$A_2(t) = 0{,}5\{1 + \cos[(\omega_1 + \omega_2)t - (\omega_1 - \omega_2)t]\} =$$
$$= 0{,}5(1 + \cos 2\omega_2 t) = \cos^2\omega_2 t$$
$$\text{и } A(t) = |\cos\omega_2 t|. \qquad (3.34)$$

Таким образом, интуитивное решение остается верным для гильбертовской огибающей, даже при небольших различиях значений частот.

4



КАК НЕ СЛЕДУЕТ ПОЯСНЯТЬ ТЕОРЕМУ КОТЕЛЬНИКОВА

4.1. Сущность теоремы

Как известно, основным содержанием теоремы Котельникова или «теоремы отсчетов» является возможность точно восстановить (интерполировать) сигнал $u(t)$ по его значениям (отсчетам), взятым в точках $t_k = k\Delta t$, $k = \ldots -2, -1, 0, 1, 2, 3, \ldots$, если спектральная плотность сигнала $S(i\omega)$ финитна, т. е. существует такое значение Ω, что $S(i\omega) = 0$ при $|\omega| > \Omega$, а $\Delta t \leqslant \pi/\Omega$. Свойства функций с финитным спектром подробно описаны и здесь повторяться не будут.

Теорема дает также способ осуществления точной интерполяции с помощью ряда Котельникова

$$u(t) = \sum_{k=-\infty}^{\infty} u(k\Delta t)\,\frac{\sin\Omega(t - k\Delta t)}{\Omega(t - k\Delta t)}. \qquad (4.1)$$

Известно (см., например, [53]), что функция с финитным спектром является целой, а целая функция не может принимать нулевых значений ни на каком интервале. Следовательно, функция с фи-

нитным спектром сама финитной быть не может. Однако часто рассматривают функцию, выражаемую рядом (4.1) при $\Delta t = \pi/\Omega$, в которой только конечное число коэффициентов $u(k\Delta t)$ может быть отличным от нуля, например коэффициенты со значениями k в пределах $k_1 \leqslant k \leqslant k_2$; такие функции рассматриваются, например, в [56]. Точки отсчета с отличными от нуля значениями в этом случае занимают интервал времени T, равный $(k_2 - k_1)\Delta t = (k_2 - k_1)\pi/\Omega = 2(k_2 - k_1)/F$, где $F = \Omega/2\pi$ — ширина спектра сигнала в обычных (не круговых) частотах.

Такой сигнал можно точно восстановить, задав значения отсчетов $u(k\Delta t)$ при $k_1 < k < k_2$. Число таких отсчетов $N = k_2 - k_1 + 1$. Выражая $k_2 - k_1$ из равенства $T = 2(k_2 - k_1)/F$, получим известное соотношение

$$N = 2FT + 1 \approx 2FT. \qquad (4.2)$$

К сожалению, иногда равенство (4.2) трактуется вульгаризованно. Еще и поныне приходится считать, что отрезок сигнала длительностью T, спектр которого лежит ниже частоты F, можно однозначно определить, задав $2FT$ значений его отсчетов через интервалы времени Δt. Это было бы верно лишь в том частном случае, когда все остальные отсчеты в моменты, кратные Δt, равны нулю, т. е. фактически тоже заданы. Заметим, что в промежутках между нулевыми отсчетами сигнал нулю не равен.

В общем же случае, как видно из (4.1), значение сигнала в любой момент t, не совпадающий с точкой отсчета, определяется всеми отсчетами и, значит, для точного восстановления сигнала необходимо знать бесконечное множество его отсчетных значений.

Правда, поскольку с увеличением $|x|$ максимумы абсолютных значений функции $\sin x/x$ довольно быстро убывают, основную роль при вычислении значения функции $u(t)$ в определенный момент времени t играют отсчеты, взятые в моменты $k\Delta t$, не очень далекие от t. Поэтому при $FT \gg 1$ ф-ла (4.2) приближенно выражает число отсчетов, существенно влияющих на значение функции $u(t)$ со спектром, лежащим ниже частоты F, на интервале длительностью T.

Сейчас это приближенное соотношение нас интересовать не будет. Сосредоточим внимание на сущности теоремы, сформулированной в начале этой главы. Заметим, что, хотя в этой формулировке для точного восстановления сигнала необходимо знать бесконечное число отсчетов, справедливость теоремы далеко не очевидна. В самом деле, речь идет в сущности о проведении кривой через измеренные в моменты $k\Delta t$ точки. Но каждому студенту хорошо известно из личного опыта, что эту кривую можно провести различными способами. Два таких варианта показаны на рис. 4.1. Теорема же говорит о точной интерполяции, т. е. о единственности такой кривой. Очевидно, что эта единственность вызвана ограниченностью спектра сигнала.

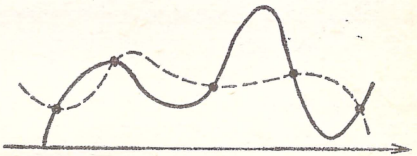


Рис. 4.1. Через заданные точки можно провести плавную кривую различным образом

4.2. Попытки наглядного пояснения теоремы

Вряд ли найдется лектор, который при изложении этого вопроса хотя бы раз в жизни не поддался искушению пояснить студентам теорему так. Поскольку спектр сигнала содержит частоты не выше F, то за время $\Delta t = 1/(2F)$, т. е. равное половине периода самой высокой частоты спектра, сигнал не может претерпеть значительных изменений. Поэтому кривая должна проходить через точки так, чтобы между ними изменяться плавно, и, как следует из (4.1), определяется по точкам отсчета однозначно.

Это весьма «правдоподобное» объяснение можно найти даже в некоторых книгах. Однако оно совершенно ошибочно. В действительности дело обстоит не так просто, о чем свидетельствует теорема, доказанная в 1957 г. Д. В. Агеевым.

Пусть на интервале (t_1, t_2) *заданы любая непрерывная функция* $u(t)$ *и произвольная частота* F. *Тогда можно построить функцию, спектр которой не содержит частот выше* F, *сколь угодно близкую (в среднеквадратичном смысле) — к* $u(t)$ *на интервале* (t_1, t_2).

Например, можно на интервале времени в 1 с задать функцию, меняющую свой знак миллион раз (скажем, задать отрезок синусоиды с частотой 1 МГц), и продолжить ее вне этого отрезка так, что спектр продолженной функции будет охватывать только область частот ниже 100 Гц, или 10 Гц, или 0,1 Гц ...

Это утверждение на первый взгляд представляется совершенно неправдоподобным. Функция с финитным спектром на интервале, меньшем полупериода высшей частоты, может, оказывается, совершать сколько угодно осцилляций. Пишущий эти

строки хорошо помнит, что, когда Д. В. Агеев доложил содержание и доказательство этой теоремы на Всесоюзной научной сессии Научно-технического общества им. А. С. Попова в Москве в 1957 г., бо́льшая часть слушателей не поверила в справедливость этой теоремы и стала искать погрешность в доказательстве.

Действительно, в это трудно было поверить людям, привыкшим связывать ширину спектра со скоростью передачи информации. Ведь можно взять отрезок широкополосного сигнала, например, с длительностью 1 с и равномерной спектральной плотностью в полосе частот до 1 МГц, который содержит до 10^6 различных независимых отсчетов и, следовательно, может нести соответственно большое количество информации, а затем в согласии с теоремой Агеева продолжить этот сигнал вне заданного отрезка так, чтобы он занимал полосу частот в 1 Гц. Тогда в канале, пропускающем частоты не выше 1 Гц, можно будет передать этот сигнал без необратимых искажений и тем самым передать за 1 с очень большое количество информации в канале с полосой пропускания 1 Гц. Все это не вязалось с привычными представлениями. К тому же изложенное Д. В. Агеевым доказательство при всей его строгости было довольно сложным и запутанным. По этой же причине мы не будем приводить это доказательство здесь. Вместо этого дадим несколько более слабую, но столь же «парадоксальную» формулировку этой теоремы, которая зато имеет весьма прозрачное доказательство. Это доказательство позволяет в то же время уяснить причину того, что совершенно верное утверждение на первый взгляд противоречит здравому смыслу. Аналогичная теорема при равномерной аппроксимации доказана в [53], с. 184—187.

4.3. Ослабленная теорема Агеева

Сформулируем следующую теорему.

Пусть на интервале (t_1, t_n) *заданы* n *точек и значения величины* $u(t)$ *в этих точках. Тогда каково бы ни было конечное число* n, *при сколь угодно малом значении* F *можно построить сколько угодно различных функций* $u(t)$, *принимающих заданные значения в указанных точках, спектры которых не содержат частот выше* F.

Прежде чем доказывать эту теорему, заметим, что она не менее парадоксальна, чем сформулированная выше. Действительно, поскольку число n не ограничено, можно задать его так, что на интервал в 1 с будет приходиться, скажем, миллион точек. Если, например, знаки значений функции в этих точках чередуются, то функция проделывает не менее миллиона осцилляций на интервале в 1 с и тем не менее может иметь спектр, лежащий ниже 1 Гц.

На первый взгляд предложенная формулировка в корне противоречит теореме Котельникова. Если n точек заданы через $\Delta t = 1/(2F)$, то согласно вульгаризованной трактовке теоремы Котельникова они должны однозначно определить функцию $u(t)$ со спектром, ограниченным частотой F. Здесь же утверждается, что таких функций может быть сколько угодно. Это противоречие, конечно, кажущееся. Сформулированная выше теорема не только не противоречит теореме Котельникова, но, как будет сейчас показано, является ее прямым следствием.

4.4. Доказательство теоремы Агеева

Обозначим u_i $(i = 1, ..., n)$ заданные значения функции $u(t)$ в заданных точках t_i. Представим искомый сигнал со спектром, занимающим полосу час-

тот ниже заданной F, в виде ряда (4.1), где $\Omega = 2\pi F$, $\Delta t = 1/(2F)$, $u(k\Delta t)$ — пока неизвестные коэффициенты ряда Котельникова.

Для упрощения доказательства предположим, что ни одна из точек t_i не совпадает с точками $k\Delta t$, $-\infty < k < \infty$. Читатель сможет без большого труда обобщить доказательство, отказавшись от этого допущения.

Выберем произвольно n целых чисел k_j и обозначим неизвестные значения искомой функции в точках $k_j\Delta t$, $j = 1, \ldots, n$ через u_j.

Будем строить функцию $u(t)$, удовлетворяющую условиям:

а) она выражается рядом (4.1) и, следовательно, ее спектр не содержит частот выше F;

б) она принимает заданные значения u_i в заданных точках t_i;

в) она равна нулю в точках $k\Delta t$ при всех k, отличных от k_j.

С учетом условия «в» ряд (4.1) вырождается в конечную сумму, содержащую n слагаемых вида $u_j \frac{\sin \Omega (t - k_j \Delta t)}{\Omega (t - k_j \Delta t)}$, в которых $u_j = u(k_j \Delta t)$ неизвестны. Для нахождения их подставим в правую часть (4.1) $t = t_i$, а в левую $u(t_i) = u_i$. Таким образом получим n линейных уравнений с n неизвестными u_j.

Коэффициентами при неизвестных являются величины $\frac{\sin \Omega (t_i - k_j \Delta t)}{\Omega (t_i - k_j \Delta t)}$. Для того чтобы эта система имела решение, необходимо и достаточно, чтобы матрица, составленная из этих коэффициентов, была невырожденной. Если это условие не выполняется, то можно, заменив хотя бы одно из произвольно выбранных значений k_j, получить невырожденную матрицу. Таким образом, построенная си-

стема определяет неизвестные u_j, подставив которые в (4.1), получим функцию, удовлетворяющую условиям «а» и «б», а следовательно, и условию теоремы.

Изменив выбранные значения k_j, можно построить сколько угодно функций, удовлетворяющих условию теоремы. Все они будут различны, поскольку их значения в точках отсчета не совпадают. Можно также построить различные функции, удовлетворяющие условию теоремы, отказавшись от условия «в» и задав вместо нулевых значений в точках отсчета $k\Delta t$, отличных от $k_j\Delta t$, любые произвольно выбранные значения. Таким образом, существует бесконечное число функций, принимающих заданные значения на конечном числе задан-

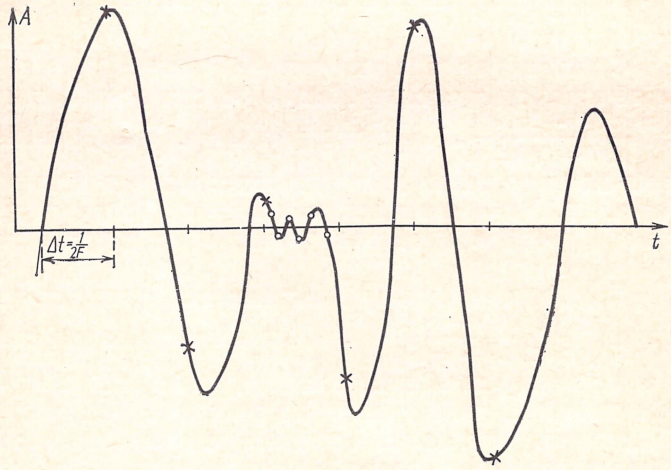


Рис. 4.2. Построение функции с финитным спектром, проходящей через произвольно заданные точки:
$n=6$, $\times$—u_j, $\circ$—u_i

ных точек, спектр которых не содержит частот выше F, что и требовалось доказать.

На рис. 4.2 показан пример функции с ограниченным спектром, проходящей через шесть заданных точек, обозначенных кружками. Вычисленные значения u_j в выбранных шести точках отсчета обозначены крестиками.

Следствие. Существует бесконечное число функций $u(t)$, спектр которых не содержит частот выше F, принимающих заданные значения u_i ($i=1, \ldots, h$) в заданных n точках отсчета $t_i=k_i\Delta t$, где $\Delta t=1/(2F)$.

Это следствие является частным случаем доказанной теоремы при специальном выборе точек t_i. В то же время оно явно свидетельствует о несостоятельности широко распространенной вульгаризованной трактовки теоремы Котельникова. Чтобы определить $u(t)$ однозначно, нужно задать $u(k\Delta t)$ при всех k, $-\infty<k<\infty$.

4.5. Можно ли передать мегабит за секунду в полосе 1 Гц?

Из доказанной теоремы следует, что отрезок сигнала длительностью в 1 с, содержащий, например, 10^6 независимых двоичных отсчетов (т. е. отсчетов, принимающих два значения, выбранных равновероятно и независимо друг от друга), можно продолжить так, что его спектр целиком уложится в полосу частот шириной в 1 Гц. Отсюда как будто бы вытекает, что по каналу, пропускающему частоты не выше 1 Гц, можно в течение секунды передать информацию в 10^6 бит.

Такой вывод верен лишь с существенными оговорками. Во-первых, следует с большой осторожностью обращаться с понятием количества информа-

ции, передаваемой в канале с ограниченной полосой пропускания. Об этом подробнее будет сказано в гл. 5.

Во-вторых, отрезок сигнала длительностью в 1 с, о котором идет речь, будучи выделен из остальной части сигнала, имеет спектр весьма широкий и поэтому сам по себе не может быть передан в рассматриваемом канале. Для того чтобы передать такой отрезок, его нельзя выделять из всего «узкополосного» сигнала, продолжающегося теоретически от $t=-\infty$ до $t=+\infty$. Практически нужно передавать узкополосный сигнал достаточно долго, прежде чем в нем сформируется отрезок длительностью в 1 с, значения которого в выбранных точках мало отличаются от заданных. Попробуем оценить необходимое время передачи.

Для этого вспомним ход доказательства теоремы, в котором определение продолженного сигнала производится путем нахождения n значений отсчетов в моменты времени, кратные Δt. Между первым и последним из этих отсчетов проходит время, не меньшее чем $n\Delta t$. Очевидно, что по меньшей мере все эти отсчеты должны быть последовательно поданы на вход канала для того, чтобы выходной сигнал содержал отрезок сигнала, несущий приблизительно[1] заданные значения в фиксированных точках. Следовательно, сигнал нужно передавать в течение времени, не меньшего $n\Delta t$.

В нашем примере $n = 10^6$ $\Delta t = 1/(2F) = 0{,}5$ с (при $F=1$ Гц) и $n\Delta t = 5 \cdot 10^5$ с (т. е. около 135 ч). Конечно, в канале с полосой пропускания 1 Гц за такое время можно, в принципе, передать 10^6 бит, так как средняя скорость передачи при этом будет

[1] Приблизительно, потому что время воздействия сигнала на вход канала конечно.

равна 2 бит/с, что соответствует так называемому пределу Найквиста [1].

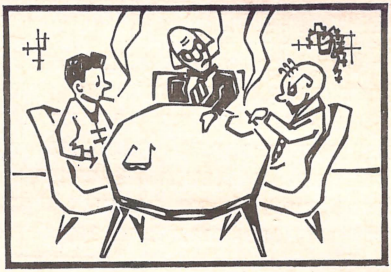
Интересно отметить еще одно обстоятельство. Если заданные значения сигнала на интервале (t_1, t_n) много раз меняют свой знак на протяжении полупериода частоты F, то средний квадрат значений u_j котельниковских отсчетов продолженного сигнала, не равных нулю, значительно больше среднего квадрата заданных значений u_i сигнала. Это легко понять, если учесть, что при большом n в формировании сигнала на заданном отрезке участвует много членов суммы (4.1), в том числе такие, для которых $\Omega(t - k\Delta t)$ достигает или превышает $2\pi n$. Очевидно, эти члены по порядку абсолютной величины должны быть близки к заданным отсчетам u_i, а поэтому хотя бы некоторые коэффициенты $u(k\Delta t)$ должны быть примерно в $2\pi n$ раз больше. Отсюда следует, что если узкополосный сигнал совершает много осцилляций на протяжении одного периода высшей частоты его спектра, то средний уровень сигнала на этом отрезке существенно ниже, чем вне его. Более точно эта особенность будет сформулирована и обоснована в гл. 6.

Таким образом, возможность передавать в узкой полосе частот большое количество информации осуществляется в данном случае путем увеличения как времени занятости канала, так и средней мощности сигнала. В этом уже ничего парадоксального нет.

---

[1] Некоторые замечания по поводу этого предела читатель найдет в гл. 5.

5



ПРЕДЕЛ ИЛИ БАРЬЕР НАЙКВИСТА

В гл. 4 был затронут вопрос о передаче информации сигналами с финитным спектром. Об этом в учебной и журнальной литературе накопилось столько взаимопротиворечивых высказываний, что даже перечислить их трудно. Для того чтобы немного разобраться в сущности противоречий, воспользуемся очень старым литературным приемом — диалогом.

Итак, послушаем беседу профессора с тремя аспирантами.

Профессор: Слышали ли вы когда-нибудь о так называемом пределе Найквиста для скорости передачи информации?

1-й аспирант: Да, я читал, что, как установил в 1928 г. Найквист, в канале с полосой пропускания F Гц за время T с не может быть передано больше чем $2FT$ бит информации. Другими словами, можно передавать информацию со скоростью не свыше 2 бит в секунду на герц.

2-й аспирант: Это не точно. За время T в полосе F можно передать $2FT$ независимых величин. Это и утверждал Найквист [31] и это вполне согласуется с теоремой Котельникова. Различие ме-

жду рассуждениями Найквиста и Котельникова, если память мне не изменяет, заключается в том, что Котельников разлагал сигнал в ряд (4.1) и определял $2FT$ отсчетных значений сигнала, а Найквист разлагал сигнал на интервале nT в ряд Фурье и определял необходимую для его восстановления полосу частот. Но и в том, и в другом случаях отсчеты сигнала могут принимать много значений. Если, например, можно надежно различать m значений отсчета, то информация в каждом отсчете составит $\log_2 m$ бит[1]. Поэтому правильнее будет сказать, что Найквист показал возможность передачи со скоростью 2 отсчета в секунду на герц, или 2 Бод на герц, если под бодом понимать единицу скорости передачи дискретных сигналов, равную одной «посылке» в секунду. Если же говорить о битах в секунду, понимая под битом двоичную единицу информации, то за время T в полосе частот F можно передать $2FT \log m$ бит. Величина m определяется, например, помехами, чувствительностью измерителя в приемном устройстве и т. д. Впрочем об этом Найквист не говорил. Он ограничивался утверждением о возможности передать $2FT$ посылок.

1-й аспирант: Я согласен с этой поправкой, но хочу ввести еще одну. Все это справедливо лишь при больших значениях произведения $2FT$. Поэтому лучше говорить не о количестве информации, переданном за время T, а о средней скорости передачи информации, достижимой при длительном использовании канала. В этом случае скорость передачи информации в канале, пропускающем частоты до F, не может превысить $2F \log m$ бит/с.

[1] В дальнейшем вместо $\log_2$ будем писать log, считая, что основание логарифмов всюду, где не оговорено противное, равно 2

3-й аспирант: Я решительно не согласен ни с вами, ни с Найквистом, если он утверждал то, что вы ему приписываете. Сигнал, прошедший через фильтр, не пропускающий частот выше F, является сигналом с финитным спектром. Такой сигнал вообще не может передавать никакой информации, на что в свое время обратил внимание Н. А. Железнов [14].

1-й аспирант: Как же так? Чем такой сигнал хуже любого другого?

3-й аспирант: Тем, что он детерминирован. Известно ведь, что сигнал с финитным спектром является целой функцией. Мы читали об этом в книге Хургина и Яковлева [53]. А целая функция бесконечно дифференцируема. Поэтому по любому небольшому отрезку такой функции можно построить ряд Тейлора и экстраполировать ее, т. е. совершенно точно предсказать все ее будущие значения. Значит, никакой новой информации эти будущие значения уже нести не могут. Скорость передачи информации оказывается равной нулю. Для того чтобы сигнал передавал новую информацию, его спектр должен отличаться от нуля на всех частотах, кроме отдельных дискретных точек.

2-й аспирант: Тут что-то не так. Я согласен, что после того, как мы приняли отрезок сигнала, проанализировали его и произвели экстраполяцию, дальнейший прием сигнала ничего нового нам не даст. Но это происходит потому, что всю информацию мы уже извлекли из принятого отрезка сигнала. Значит, проанализировав конечный отрезок сигнала с финитным спектром, можно получить информацию обо всем бесконечном сигнале, характеризуемом бесконечным числом независимых отсчетов. Следовательно, скорость передачи информации окажется не нулевой, а бесконечной.

1-й аспирант: Час от часу не легче. То скорость была нулевой, то она вдруг стала бесконечной! Мне совершенно ясно, что бесконечной она быть не может, так как это, во-первых, противоречит пределу Найквиста, с которого мы начали...

2-й аспирант: Вовсе нет. Мы ведь установили, что по Найквисту скорость передачи информации ограничена величиной $2F \log m$. Но поскольку никаких помех мы не учитываем и полагаем, что принятый сигнал измеряется абсолютно точно, то $m=\infty$ и отсюда следует, что скорость передачи ничем не ограничена.

1-й аспирант: С этим, пожалуй, можно согласиться. Но бесконечная скорость противоречит и принципу относительности, согласно которому скорость передачи сигнала ни в каких условиях не может превзойти скорость света[1].

Профессор: Тут уж я должен вмешаться. Не следует путать скорость передачи сигнала по каналу в смысле количества каких-то элементов (посылок, носителей информации), передаваемых в единицу времени, имеющую размерность 1/с, и скорость прохождения сигнала в канале, в смысле длины пути, пройденного сигналом за единицу времени, имеющую размерность метр в секунду. Последняя, конечно, не может превысить скорость света. Мы же здесь говорим о скорости передачи в первом смысле, которая ничего общего со скоростью света иметь не может, хотя бы потому, что имеет совершенно другую размерность. Но я жалею, что перебил вас. Разгоревшаяся между вами дискуссия очень интересна, и мне бы хотелось, что-

[1] Этот «довод» не выдуман автором. Он был выдвинут одним из участников дискуссии на научной конференции в конце 50-х годов.

бы вы сами довели ее до той истины, которая рождается в спорах.

1-й аспирант: Прошу прощения за то, что аппелировал к не относящейся к делу скорости света. Сейчас мне кажется, что вы оба правы. Когда начинается передача сигнала с финитным спектром, из самого начального отрезка можно выделить всю информацию, содержащуюся во всем будущем бесконечном сигнале. В этом отрезке времени скорость передачи информации бесконечна. Во все же остальное время скорость передачи равна нулю, так как сигнал ничего нового для получателя уже не содержит. Интересно бы вычислить, исходя из этого, среднюю скорость передачи информации.

3-й аспирант: С таким компромиссным решением я согласиться не могу. Представьте, например, что вам нужно передавать информацию о футбольном матче и вам для этого предоставили канал со строго финитной полосой пропускания. Сформированный вами сигнал детерминирован. Как же вы сможете ввести в него информацию о неожиданно забитом голе? Может быть она уже заложена в самом начальном отрезке сигнала? Значит, проанализировав его, можно заранее предсказать исход матча и его ход со всеми подробностями. Поскольку это невозможно, скорость передачи информации по такому каналу все время равна нулю.

2-й аспирант: Вероятно здесь нужно учесть групповое запаздывание сигнала в канале. В идеальном канале со строго финитной полосой пропускания это запаздывание бесконечно. Но не будем требовать идеальности канала. Если канал имеет амплитудно-частотную характеристику, очень близкую к П-образной и физически реализуемую, то время группового запаздывания в нем должно быть огромным, скажем, должно превышать длитель-

ность матча. Значит, сигнал с почти финитным спектром сформируется на выходе канала в то время, когда матч уже закончился и весь репортаж о нем уже произнесен перед микрофоном. Помехи мы считаем отсутствующими. Тогда ничего странного не будет в том, что, проанализировав самое начало сформировавшегося сигнала, можно выявить из него всю передававшуюся информацию. Это подобно тому, как по небольшой части голограммы можно восстановить все заложенное в ней изображение.

1-й аспирант: Мне не нравится, что мы все время пренебрегаем помехами. Что изменится, если мы их учтем?

3-й аспирант: Абсолютно ничего, если помехи тоже прошли через наш канал и имеют финитный спектр в той же полосе, что и сигнал. Значит, сумма сигнала и помехи является целой функцией, которую можно экстраполировать с любой точностью, и, следовательно, она не может быть носителем информации.

2-й аспирант: Почему же помехи обязательно должны пройти через тот же канал? Ими могут быть тепловые шумы приемника или, если фильтр является частью приемника, тепловые шумы элементов аппаратуры, включенных после фильтра. Тогда спектр суммы сигнала и помехи не будет финитным. Точная экстраполяция окажется невозможной, и скорость передачи все время будет ненулевой.

1-й аспирант: Смотрите же, до чего мы договорились! Если сигнал, прошедший через канал со строго финитной полосой пропускания, поступает на приемник без помех, то скорость передачи информации равна нулю. Если же подмешать к нему шум, то скорость передачи информации становит-

ся ненулевой! Согласитесь, что это нелепый результат — добавление шума не может увеличить скорость передачи информации.

2-й аспирант: Почему же? Я могу привести другой очень простой пример с таким же результатом.

Предположим, что для передачи сообщений по некоторому дискретному каналу используется система, в которой каждый кодовый блок повторяется 15 раз подряд и решение принимается по мажоритарному принципу, т. е. по большинству принятых за 15 раз значений каждого символа. Пусть вначале помех в канале нет. Тогда при первой же передаче кодовый блок принимается без ошибок. Следовательно, остальные 14 повторений того же блока не несут информации и скорость передачи после того, как первый блок закончен, равна нулю. Но если только в канале имеется шум, то при первой передаче блок может быть принят с ошибками и тогда все его повторения несут некоторую информацию, так как они уже не известны априори.

1-й аспирант: Пожалуй это так. Добавление помехи может в некоторых случаях увеличить скорость передачи информации на некотором отрезке времени. Но оно, по-видимому, не может увеличить средней скорости передачи информации за все время использования канала. Если повторять 15 раз кодовый блок в канале без помех, то средняя скорость передачи информации будет равна $I(X, X)/15T_0$, где $I(X, X)$ — количество информации, содержащееся в передаваемом блоке, T_0 — длительность однократной передачи блока. Если же в канале имеются помехи, то средняя скорость передачи информации равна $I(Y, X)/15T_0$, где $I(Y, X)$ — количество информации, содержащееся в принятых блоках Y относительно переданного блока X. Собственно говоря, эта формула справедлива и

в отсутствие помех, но в этом случае $Y=X$. Известно, что $I(Y, X) \leqslant I(X, X)$. Поэтому добавление помехи не может увеличить среднюю скорость передачи информации. По-видимому, то же самое имеет место и в канале с финитной полосой пропускания. Там очень большое количество информации передается в первый же момент, вслед за чем скорость передачи информации падает до нуля. Но средняя скорость за все время функционирования канала остается, вообще говоря, конечной. Я уже говорил, что предел Найквиста относится только к средней скорости — она не может превысить $2F \log m$.

2-й и 3-й аспиранты: Да, это похоже на истину. Но все же многое еще осталось неясным.

Профессор: Совершенно с вами согласен. В вашем споре возник пока что зародыш истины. Для полного его развития нужно еще многое додумать. Я не стану этого делать за вас, но попытаюсь дать вам некоторые наводящие идеи.

Прежде всего сама постановка вопроса не вполне корректна. Цепь с финитной амплитудно-частотной характеристикой физически не реализуема. Следовательно, такого канала в природе существовать не может. А поэтому незаконно ставить вопрос о его пропускной способности. Как показали Пэйли и Винер [32], амплитудно-частотная характеристика $K(\omega)$ физически реализуемой цепи должна удовлетворять условию

$$\int_0^\infty \frac{|\ln K(\omega)|}{1+\omega^2} d\omega < \infty. \qquad (5.1)$$

Это условие не выполняется, если $K(\omega)=0$ хотя бы на конечном интервале. Оно не удовлетворяется и в том случае, если при больших ω величина $K(\omega)$

уменьшается быстрее, чем некоторая степень ω. Следовательно, таких каналов быть не может и о скорости передачи информации в них говорить не следует.

Конечно, можно рассматривать последовательность каналов K_1, K_2, ..., амплитудно-частотные характеристики которых стремятся в некотором смысле (например, в смысле равномерной сходимости) к П-образной характеристике идеального фильтра, и отыскивать предел их пропускной способности. Однако этот предел ничему реальному не соответствует, так как время задержки в канале K_n стремится вместе с n к бесконечности. Уже поэтому можно утверждать, что по такому «идеальному» каналу информацию передавать нельзя.

1-й аспирант: Хорошо, не будем говорить о канале с финитной полосой пропускания. Пусть канал пропускает без искажений и задержек все частоты. Но сигнал-то может иметь финитный спектр. Мы можем его сформировать, задавшись некоторым конечным числом «котельниковских» отсчетов u_k, $k=1$, ..., n, положив все остальные отсчеты равными нулю, подставив их в ряд

$$u(t) = \sum_{k=1}^{n} u_k \frac{\sin \Omega (t - k\Delta t)}{\Omega (t - k\Delta t)}, \qquad (5.2)$$

$$\Delta t = \pi/\Omega,$$

и подавая на вход канала полученные значения $u(t)$, скажем, от $t=0$ до $t=n\Delta t$. Этот сигнал имеет финитный спектр. Передать его отрезок можно без всякой задержки. Мы можем вложить в него при формировании информацию, выбирая значения u_k. Следовательно, отрезок сигнала с финитным спектром может передавать информацию со скоростью, определяемой пределом Найквиста.

3-й аспирант: Это не верно. Если вырезать отрезок функции с финитным спектром на интервале (t_1, t_2) и положить ее равной нулю вне этого интервала, то спектр полученного отрезка уже не финитный. Поэтому-то и понадобился канал с неограниченной полосой пропускания. Следовательно, сформированный отрезок сигнала не является детерминированным (хотя бы потому, что он может закончиться в любой момент). С другой стороны, пусть получателю сообщения заранее известно, что передаваемый сигнал выражается суммой (6.2), и неизвестны только значения коэффициентов u_k. Тогда, если помех в канале нет, вовсе не обязательно передавать отрезок этого сигнала от $t=0$ до $t=h\Delta t$. Можно ограничиться передачей сколь угодно короткого отрезка сигнала и даже не обязательно лежащего внутри интервала $(0, n\Delta t)$. Достаточно даже передать ненулевые значения $u_i=u(t_i)$ в любых n различных точках t_i, так как по ним можно восстановить все u_k.

2-й аспирант: Мне кажется, что мы пришли к тривиальному результату. Формируя сигнал $u(t)$ (6.2), мы ставили перед собой цель передать информацию о последовательности n действительных чисел u_k. Далее оказалось, что вместо сигнала $u(t)$ можно передать n других чисел — значений $u(t_i)$ в n точках, не обязательно совпадающих с точками $k\Delta t$. Это совершенно естественно, так как ф-ла (6.2) определяет взаимно однозначное соответствие между совокупностями n чисел u_k и n других чисел u_i. А отрезок сигнала с финитным спектром здесь вовсе не при чем.

Профессор: Давайте подведем итог. Прежде всего я хочу вам посоветовать не терять чувства исторической перспективы. Работа Найквиста [31] появилась задолго до того, как возникло современ-

ное понятие пропускной способности канала. В работе Найквиста даже термина «пропускная способность» нет, он появился впервые, да и то не в современном смысле, в статье Котельникова в 1933 г.

Интересно все же отметить, что у Найквиста уже применяются понятия информация и избыточность (redundance), причем в них вкладывается смысл, очень близкий к современному. Вот перевод основных тезисов статьи Найквиста:

«1. Требуемая полоса частот прямо пропорциональна скорости передачи (signaling speed).

2. Повторяемый телеграфный сигнал [1] можно рассматривать как состоящий из синусоидальных компонент. Если амплитуду и фазу или действительную и мнимую части этих компонент отложить по оси ординат, а частоты — по оси абсцисс и ось частот разделить на части, каждая из которых представляет полосу частот, численно равную скорости передачи [2], то оказывается, что информация, содержащаяся в этих полосах, идентична, и можно сказать, что эти полосы взаимно избыточны.

3. Минимальная полоса, требуемая для точной интерпретации сигнала, численно равна скорости передачи и не зависит от числа используемых значений тока».

Отсюда видно, что Найквист хорошо понимал возможность увеличения количества передаваемой

---

[1] Речь идет о периодическом повторении отрезка сигнала, состоящего из n элементов, для пояснения разложения в ряд Фурье.

[2] Далее Найквист уточняет, что он понимает под скоростью передачи: «Скорость передачи S обычно обозначается числом точек, переданных в секунду, и определяется как число элементов сигнала в секунду, разделенное на 2». Таким образом, величина S вдвое меньше той величины, которую в настоящее время называют технической скоростью передачи.

информации путем перехода от двоичных посылок к m-ичным. Он устанавливал только предел для количества посылок, передаваемых в секунду.

Конечно, говорить о пропускной способности канала без учета помех, вообще говоря, бессмысленно. Однако иногда удобно вместо непосредственного учета помех наложить ограничения на число значений принимаемых сигналом в определенные моменты времени. Пусть, например, сигнал может принимать в моменты времени $t_k = k\Delta t$, $k = 1, \ldots, n$ лишь одно из двух значений: $+1$ или -1, а в моменты времени $k\Delta t$ при $k<1$ или $k>n$ он равен нулю. Такой сигнал может иметь финитный спектр с верхней частотой $F=1/2\Delta t$ и в этом случае представляется суммой (5.2). Количество информации в этом отрезке сигнала на интервале $T=n\Delta t$, очевидно, равно $n=2FT$ бит. Это иногда и называют «пределом Найквиста». Однако слово «предел» здесь не очень уместно. Во-первых, как понимал уже сам Найквист, это количество информации можно увеличить в $\log m$ раз, если различать в моменты времени $k\Delta t$ не два, а m уровней сигнала. Во-вторых, как мы уже видели, $2FT$ бит информации можно передать отрезком сигнала (5.2) длительностью $T_1 \ll T$. Правда, при этом нужно в каких-либо n точках различать не два уровня сигнала, а значительно больше, тем больше, чем меньше T_1/T. Практически это возможно лишь при ничтожном уровне аддитивных помех.

Как бы то ни было, отсюда следует, что термин «предел Найквиста» неудачен, он не устанавливает никаких строгих предельных соотношений, аналогичных, например, пропускной способности Шеннона или потенциальной помехоустойчивости Котельникова.

1-й аспирант: Так не лучше ли вообще отказаться от понятия «предел Найквиста»?

Профессор: Думаю, что полностью от него отказываться не стоит, и вот по какой причине. В настоящее время разработано много различных систем передачи дискретных сообщений для разнообразных каналов, в том числе для таких, у которых амплитудно-частотная характеристика довольно быстро падает за пределами полосы частот F. Ее условно можно назвать полосой пропускания канала. Скорость передачи $2F$ бит/с также условно можно назвать «скоростью Найквиста». Среди различных существующих систем связи имеются и работающие со скоростью, большей найквистской. Но вот что примечательно. Все более или менее простые системы в каналах с различными помехами обеспечивают скорость передачи не более 0,3—0,5 от найквистской. Для достижения скорости Найквиста систему приходится заметно усложнять, а превышение этой скорости требует еще значительно большего усложнения.

Применяя грубую аналогию, можно сказать, что скорость Найквиста в теории связи играет не такую роль, как скорость света в физике, а скорее такую, как скорость звука («звуковой барьер») в авиации. Достигнуть ее и превзойти можно, но это требует значительных усилий.

О причинах трудной преодолимости «барьера Найквиста» можно было бы говорить много. Остановимся лишь на одном поясняющем примере.

Рассмотрим канал с отношением мощности сигнала к мощности аддитивного гауссовского шума на выходе канала $P_с/P_ш$ при передаче сигналов, занимающих условную полосу частот F. Пропускная способность такого канала при равномерном спектре шума равна по Шеннону

$$C = F \log (1 + P_c/P_ш). \qquad (5.3)$$

Известно, что при приближении скорости передачи к пропускной способности канала требуется резко усложнять кодирование.

Для приближенной оценки положим, что граница между приемлемой и чрезмерной сложностью кодирования имеет место при скорости передачи $R = C/2$. Тогда для относительно простых систем

$$R \leqslant 0{,}5 F \log (1 + P_c/P_ш) = 0{,}25 R_N \log (1 + P_c/P_ш),$$

где $R_N = 2F$ — «скорость Найквиста» (условная).

Отсюда легко получить

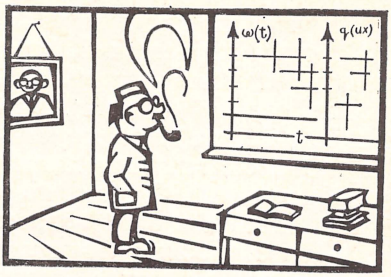
$$P_c/P_ш \geqslant 2^{4R/R_N} - 1. \qquad (5.4)$$

Если необходимую мощность сигнала при скорости, равной четверти найквистовской, принять за единицу, то для достижения половины найквистовской скорости потребуется сигнал с мощностью 3. Для того же, чтобы добиться полной найквистовской скорости, нужно увеличить мощность сигнала до 15. Дальнейшее повышение скорости за пределом «барьера Найквиста» потребует еще более резкого увеличения мощности сигнала. Так, для превышения «барьера» в 2 раза мощность сигнала должна быть в 255 раз выше, чем при $R = R_N/4$.

Если надлежащий запас мощности сигнала имеется, то «барьер» можно преодолевать различными способами. Так, можно передавать «посылки» с найквистовской скоростью $2F$ отсчетов в секунду, но увеличить число уровней m. Можно, наоборот, сохранить небольшое число уровней посылок, но передавать их чаще, чем $2F$ в секунду. При этом возникает межсимвольная интерференция, но она,

в принципе, не препятствует извлечению информации из принятого сигнала, если уровень аддитивного шума мал. Возможны и другие методы передачи.

Этот вопрос приобретает практическую актуальность, если необходимо передавать большие потоки информации по каналу, имеющему ограниченную полосу пропускания при низком уровне аддитивных помех и малом затухании сигнала. Конкретная система связи, работающая выше «барьера Найквиста», должна выбираться с учетом всех особенностей канала, таких, как мультипликативные помехи (замирания), многолучевое распространение сигнала и т. д.



6

ЕЩЕ О СПЕКТРЕ СИГНАЛА И ЕГО МГНОВЕННОЙ ЧАСТОТЕ

6.1. Сущность вопроса

Выше говорилось о различии между мгновенной частотой сигнала и частотами его спектральных составляющих. Можно ли, однако, утверждать, что между мгновенной частотой и спектром сигнала нет никакой связи? Например, может ли существовать сигнал, спектр которого сосредоточен в пределах от 100 до 300 Гц, а мгновенная частота принимает значения от 400 до 500 Гц?

Вероятно, каждый инженер ответит на этот вопрос отрицательно. Тем не менее, вряд ли многие из них сумеют точно сформулировать сущность связи между спектром сигнала и зависимостью его мгновенной частоты от времени.

Этот вопрос заинтересовал автора уже очень давно, во всяком случае раньше 1954 г. Интерес к свойствам мгновенной частоты не был вызван одной любознательностью — он был тесно связан с задачами, решавшимися в проводившейся тогда автором разработке. Тем не менее, разработка была успешно завершена, а ясности в поведении мгно-

венной частоты так и не удалось установить. С тех пор прошло уже около 25 лет, и все эти годы автор продолжал между делом размышлять о связи между мгновенной частотой и спектром. Основные из полученных при этом результатов опубликованы в [45]. Они оказались полезными также для теории линейных цепей. Однако, по-видимому, еще многое можно сделать в этой области, в частности отыскать более удобные формулировки уже известных соотношений. Основное содержание данной главы — рассказ о поисках автора в этом направлении.

6.2. Некоторые заблуждения

Первая мысль, которая приходит в голову, заключается в том, что область изменения мгновенной частоты должна совпадать с полосой частот, занимаемой спектром сигнала. Назовем эту мысль *предположением А*. Оно явно ошибочно. На нем базировались предложения об использовании узкополосной ЧМ для размещения большого числа сигналов в узком диапазоне частот (см. гл. 2). Тем не менее, в различных видоизмененных формах эту мысль продолжают проповедовать и сейчас. Ошибочность такого предположения проще всего демонстрируется на примере АМ сигнала $A(t)\cos\omega_0 t$ с постоянной мгновенной частотой ω_0. Как известно, его спектральная плотность не сосредоточена на частотах $\pm\omega_0$, а содержит также боковые полосы, которые при достаточно широком спектре огибающей $A(t)$ могут простираться очень далеко от ω_0. Если постоянная составляющая $A(t)$ равна нулю, то $A(t)\cos\omega_0 t$ представляет сигнал с балансной модуляцией и вовсе не содержит составляющей на несущей частоте ω_0.

Итак, область частот, содержавшихся в спектре сигнала, и множество значений, пробегаемых мгновенной частотой, не совпадают.

Легко опровергнуть также следующие два приходящие в голову взаимоисключающие предположения.

Предположение B: спектр сигнала занимает полосу частот, не превышающую область изменения мгновенной частоты.

Предположение C: мгновенная частота может принимать только значения, не выходящие за пределы полосы частот, занятой спектром сигнала.

Предположение B сразу опровергается тем же примером амплитудномодулированного сигнала $A(t)\cos\omega_0 t$. Если спектр $A(t)$ лежит целиком ниже частоты ω_0, то мгновенная частота такого сигнала постоянна и равна ω_0, тогда как спектр содержит боковые полосы и имеет ширину, равную удвоенной ширине полосы спектра огибающей $A(t)$.

Предположение C опровергается другим примером — суммой двух гармонических функций с близкими значениями частот и амплитуд. Пусть $u(t) = a_1\cos\omega_1 t + a_2\cos\omega_2 t$. Спектр этого сигнала дискретный и содержит две составляющие. Сопряженным по Гильберту сигналом является

$$\hat{u}(t) = a_1\sin\omega_1 t + a_2\sin\omega_2 t.$$

Мгновенная частота определяется ф-лой (3.9)

$$\omega(t) = \frac{\overset{\wedge}{u'}(t)\,u(t) - u'(t)\,\hat{u}(t)}{u^2(t) + \hat{u}{}^2 t} =$$

$$= \frac{\omega_1 a_1^2 + \omega_2 a_2^2 + (\omega_1 + \omega_2)\,a_1 a_2\cos(\omega_2 - \omega_1)\,t}{a_1^2 + a_2^2 + 2a_1 a_2\cos(\omega_2 - \omega_1)\,t}. \qquad (6.1)$$

и, следовательно, является периодической функцией с периодом $2\pi/(\omega_2 - \omega_1)$.

Пусть для определенности $a_2 < a_1$. Обозначим $a_2 = a_1(1-\varepsilon)$, $\omega_2 - \omega_1 = \Delta\omega$. Тогда

$$\omega(t) = \omega_1\left[1 + \frac{\Delta\omega}{\omega_1}\,\frac{(1-\varepsilon)^2 + (1-\varepsilon)\cos\Delta\omega t}{1 + (1-\varepsilon)^2 + 2(1-\varepsilon)\cos\Delta\omega t}\right]. \quad (6.2)$$

Если $\Delta\omega \ll \omega_1$, то второй член в квадратных скобках на первый взгляд значительно меньше единицы, так что мгновенная частота близка к частоте ω_1 той спектральной составляющей, у которой больше амплитуда. Это можно проверить для крайних случаев. Если $\varepsilon = 0$, то второй член равен $\frac{\Delta\omega}{2\omega_1}$, так что $\omega(t) = \omega_1 + \Delta\omega/2 = \frac{\omega_1 + \omega_2}{2}$, как и можно было ожидать из соображений симметрии. Если же $\varepsilon = 1$, то второй член равен 0 и $\omega(t) = \omega_1$. Однако не будем торопиться. Рассмотрим промежуточный случай, когда $0 < \varepsilon \ll 1$. На рис. 6.1 представлена функция $\omega(t)$ для трех значений $\varepsilon = 0{,}8$; 0,5 и 0,1 при $\Delta\omega/\omega_1 = 0{,}2$. При очень малых ε мгновенная частота отклоняется от ω_1 значительно больше, чем на величину $\Delta\omega$, которая представляет собой ширину спектра.

Действительно, в моменты времени, когда $\Delta\omega t$ принимает значения $(2k+1)\pi$, из (6.2) имеем

$$\omega(t) = \omega_1 + \Delta\omega(\varepsilon^2 - \varepsilon)/\varepsilon^2 = \omega_1 + \Delta\omega(1 - 1/\varepsilon) = \\ = \omega_2 - \Delta\omega/\varepsilon. \quad (6.3)$$

При достаточно малых значениях ε мгновенная частота может стать сколь угодно большой (если $\Delta\omega < 0$) или сколь угодно малой и даже отрицательной[1] (если $\Delta\omega > 0$).

---

[1] В отличие от некоторых других работ, мы называем мгновенной частотой $\frac{d\Phi}{dt}$, а не $\left|\frac{d\Phi}{dt}\right|$ и не $\frac{d\Phi}{dt} - \omega_0$.

Кривые эти выглядят довольно парадоксально. Самое странное то, что резкие отклонения мгновенной частоты происходят в сторону, противоположную расстройке $\Delta\omega$ более слабой составляющей. Однако они вполне реальны — их можно наблюдать на экране осциллографа, подключенного к выходу частотного дискриминатора, если на вход его ограничителя подать сумму двух гармонических сигналов.

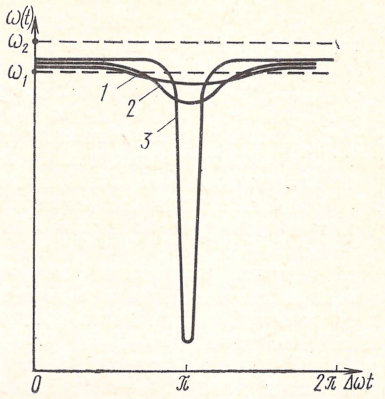


Рис. 6.1. Мгновенная частота суммы двух гармонических сигналов:
1—$\varepsilon=0{,}8$ $(a_2/a_1=0{,}2)$; *2*—$\varepsilon=0{,}5$ $(a_2/a_1=0{,}5)$; *3*—$\varepsilon=0{,}1$ $(a_2/a_1=0{,}9)$

6.3. Поиски

Итак, ширина спектра может быть сколь угодно у́же и сколь угодно шире области изменения мгновенной частоты или, как говорят, удвоенной девиации, понимая под девиацией предел отклонения мгновенной частоты от среднего значения в одну сторону. Поэтому связь между значениями мгновенной частоты и частотами спектра должна быть не столь однозначной.

Рассматривая различные другие примеры, автор обнаружил, что чем больше отклоняется мгновенная частота от своего среднего значения, тем быстрее она возвращается обратно в область значений, близких к среднему. Кроме того, эти средние значения, как показывает опыт, близки к средней

частоте спектра. Эти наблюдения позволили автору выдвинуть в 1955 г. следующее предположение (назовем его D):

Мгновенная частота сигнала может принимать любые значения, однако неравенство $\omega(t) > 2\pi(f_{\text{макс}} + \Delta f)$ или $\omega(t) < 2\pi(f_{\text{мин}} - \Delta f)$, где $f_{\text{мин}}$ и $f_{\text{макс}}$ — нижняя и верхняя частоты спектра, может выполняться только в течение промежутка времени, не превышающего $1/\Delta f$.

Это предположение было проверено на большом числе примеров. Однако попытка строгого доказательства удалась лишь при некоторых дополнительных условиях, наложенных на сигнал.

Продолжая размышлять над этим вопросом, естественно было попытаться заменить точную ширину спектра, имеющую смысл только для финитного спектра, шириной некоторой условной полосы частот. Среди различных определений такой полосы (например, полоса, в которой сосредоточено 99% мощности сигнала, или полоса, за пределами которой амплитуды дискретных составляющих спектра не превышают заданной величины, ширина прямоугольника, равновеликого фигуре, образуемой спектральной плотностью мощности, и т. д.) наиболее привлекательным является определение средней квадратичной ширины спектра, по-видимому, впервые предложенное С. М. Рытовым [35], используемое многими авторами, в частности В. И. Бунимовичем [4], и представляющее собой второй центральный момент положительной ветви нормированной энергетической спектральной плотности. Эта привлекательность вызвана, во-первых, тем, что в определение этой полосы не входит никаких произвольных величин (вроде 99% мощности либо высоты эквивалентного прямоугольника), во-вторых, тем, что оно применимо практически для всех сигналов, как детерминированных, так и случайных (стационарных), в-третьих, тем, что моменты уже широко применялись для решения многих вопросов теории связи.

Будем обозначать частоты спектра (аргумент спектральной плотности) ω, а нормированную энергетическую спектральную плотность — $g(\omega)$. Напомним, что для детерминированных сигналов

$$g(\omega) = \frac{1}{2\pi} \lim_{T \to \infty} \frac{\left| \int_{-T}^{T} u(t)\, e^{-i\omega t}\, dt \right|^2}{\int_{-T}^{T} u^2(t)\, dt}, \qquad (6.4)$$

в частном случае для сигналов, интегрируемых с квадратом, эта формула дает

$$g(\omega) = |S(i\omega)|^2/E, \tag{6.5}$$

где $S(i\omega) \int_{-\infty}^{\infty} u(t)e^{-i\omega t}\,dt$ — обычная спектральная плотность сигнала; $E = \int_{-\infty}^{\infty} u^2(t)dt$ — его энергия.

Для сигналов с конечной мощностью из общего определения следует

$$g(\omega) = \frac{1}{P}\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} u(t)\,e^{-i\omega t}\,dt, \tag{6.6}$$

где

$$P = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} u^2(t)\,dt \tag{6.7}$$

— средняя мощность сигнала.

Для случайных стационарных сигналов в соответствии с теоремой Винера — Хинчина

$$\left.\begin{aligned} g(\omega) &= \frac{1}{\pi}\int_0^{\infty} R(\tau)\cos\omega\tau\,d\tau, \\ R(\tau) &= 2\int_0^{\infty} g(\omega)\cos\omega\tau\,d\omega, \end{aligned}\right\} \tag{6.8}$$

где $R(\tau)$ — нормированная корреляционная функция (или коэффициент корреляции) процесса $u(t)$ (см., например, [26]). При таком определении всегда $g(\omega) = g(-\omega)$ и $\int_{-\infty}^{\infty} g(\omega)\,d\omega = 1$.

Средней частотой спектра называется величина (первый начальный момент функции $g(|\omega|)$

$$\omega_I = \int_{-\infty}^{\infty} |\omega| g(\omega) d\omega = 2\int_0^{\infty} \omega g(\omega) d\omega, \qquad (6.9)$$

если этот интеграл существует[1]).

Средний квадрат частоты спектра определяется как второй начальный момент $g(\omega)$:

$$\omega_{II}^2 = \int_{-\infty}^{\infty} \omega^2 g(\omega) d\omega = 2\int_0^{\infty} \omega^2 g(\omega) d\omega, \qquad (6.10)$$

а *средний квадрат ширины спектра* — как второй центральный момент

$$(\delta\omega)^2 = \int_{-\infty}^{\infty} (|\omega| - \omega_I)^2 g(\omega) d\omega = \omega_{II}^2 - \omega_I^2. \qquad (6.11)$$

Здесь также предполагается, что все эти интегралы существуют.

В отличие от спектральных частот ω, мгновенную частоту будем обозначать Ω. Она является функцией времени и может характеризоваться средним значением по времени:

$$\Omega_{ср} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \Omega(t) dt. \qquad (6.12)$$

При попытках найти связь между числовыми характеристиками спектра и мгновенной частоты возникло новое *предположение Е:* средняя мгновенная частота равна средней частоте спектра ω_I.

---

[1]) Более общее определение ω_I и других вводимых ниже моментов, когда соответствующие интегралы расходятся, приведено в [4].

Известно, что это равенство выполняется для стационарного нормального процесса с нулевым средним значением и симметричным спектром[1]. Далее легко убедиться, что это же равенство выполняется для любых сигналов с чисто амплитудной модуляцией вида

$$u(t) = A[1 + mx(t)]\cos(\omega_0 t + \varphi), \qquad (6.13)$$

у которых $\Omega(t) = \omega_0 = \text{const} = \Omega_{\text{ср}}$. Это же справедливо и при балансной модуляции (АМ без несущей), когда $u(t) = x(t)\cos(\omega_0 t + \varphi)$, где $x(t)$ — любой модулирующий сигнал с нулевым средним значением (т. е. без постоянной составляющей), причем предполагается, что весь спектр $x(t)$ расположен ниже частоты ω_0[2]. Следующим рассмотренным примером был сигнал с чисто угловой модуляцией $u(t) = A\cos\Phi(t)$, где $A = \text{const}$, $\Phi(t)$ — произвольная функция, $\Omega(t) = \dfrac{d\Phi}{dt}$. Этот пример также выдержал проверку, для него $\omega_I = \Omega_{\text{ср}}$.

После стольких проверок предположение E представлялось весьма правдоподобным и автор принялся отыскивать для него доказательство, справедливое в общем случае или хотя бы при несильно ограничивающих дополнительных условиях. Однако доказательство никак не клеилось. И тог-

---

[1] Для такого процесса верно также следующее соотношение:

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}|\Omega(t)|\,dt = \omega_{II}.$$

[2] В противном случае $|x(t)|$ не является огибающей в смысле определения, основанного на преобразовании Гильберта, а следовательно, ω_0 не является постоянной мгновенной частотой, что затрудняет проверку равенства $\Omega_{\text{ср}} = \omega_I$.

да автор вспомнил мудрый совет, сформулированный одним из его учеников: если некоторое предположение кажется почти очевидным, но строго доказать его никак не удается, то скорее всего оно не верно.

Действительно, рассмотрев еще несколько более сложных случаев, удалось найти опровергающий пример.

6.4. Находки

После многих попыток найти более универсальное предположение о связи между численными характеристиками спектра и мгновенной частотой возникла идея использовать огибающую сигнала в качестве весового множителя при определении средних значений мгновенной частоты. Таким образом, можно ввести понятие *средней (взвешенной) мгновенной частоты*

$$\Omega_I = \lim_{T\to\infty} \frac{\int_{-T}^{T} A^2(t)\,\Omega(t)\,dt}{\int_{-T}^{T} A^2(t)\,dt}, \qquad (6.14)$$

среднего (взвешенного) квадрата мгновенной частоты

$$\Omega_{II}^2 = \lim_{T\to\infty} \frac{\int_{-T}^{T} A^2(t)\,\Omega^2(t)\,dt}{\int_{-T}^{T} A^2(t)\,dt} \qquad (6.15)$$

и среднего (взвешенного) квадрата девиации

$$(\delta\Omega)^2 = \lim_{T\to\infty} \frac{\int_{-T}^{T} A^2(t)\,[\Omega(t) - \Omega_I]^2\,dt}{\int_{-T}^{T} A^2(t)\,dt} = \Omega_{II}^2 - \Omega_I^2. \qquad (6.16)$$

После введения этих величин все пошло очень гладко. Сразу удалось доказать *теорему 1:*

$$\omega_I = \Omega_I, \qquad (6.17)$$

а затем и *теорему 2:*

$$\omega_{II}^2 = \Omega_{II}^2 + \lim_{T\to\infty} \frac{\int_{-T}^{T} A'^2(t)\,dt}{\int_{-T}^{T} A^2(t)\,dt}. \qquad (6.18)$$

Очевидным следствием этих теорем является связь между средним квадратом девиации и средним квадратом ширины спектра:

$$(\delta\omega)^2 = (\delta\Omega)^2 + \lim_{T\to\infty} \left(\int_{-T}^{T} A'^2(t)\,dt \Big/ \int_{-T}^{T} A^2(t)\,dt \right). \qquad (6.19)$$

Обе эти теоремы справедливы, если фигурирующие в них интегралы существуют, а сигнал удовлетворяет некоторым условиям, практически не сужающим область применения. Они подробно перечислены в [45], где также даны строгие доказательства этих теорем. Для стационарных случайных процессов эти теоремы верны и при замене усреднения мгновенной частоты по времени усреднением по ансамблю. В этом случае

$$\left.\begin{aligned} \Omega_I &= (1/2\sigma^2) < A^2\,\Omega >, \\ \Omega_{II}^2 &= (1/2\sigma^2) < A^2\,\Omega_t^2 >, \end{aligned}\right\} \qquad (6.20)$$

где σ^2 — дисперсия случайного сигнала, угловые скобки обозначают математическое ожидание.

В [48] дано более простое доказательство теоремы 1, применимое, однако, лишь к детерминированным сигналам из L_2 (т. е. с интегрируемым квадратом).

Остановимся на выводах, вытекающих из упомянутых теорем. Во-первых, они свидетельствуют о наличии глубокой связи между спектром сигнала и его мгновенной частотой. Средняя взвешенная мгновенная частота всегда в точности совпадает со средней частотой спектра. Из (6.14) и (6.19) легко видеть, в каких случаях должно выполняться предположение E, в соответствии с которым $\Omega_{ср} = \omega_I$. Очевидно, для этого необходимо и достаточно выполнение равенства $\Omega_I = \Omega_{ср}$ или

$$<A^2(t)\,\Omega(t)>/<A^2(t)> = <\Omega(t)>,$$

где угловые скобки обозначают усреднение либо по времени (для детерминированной реализации сигнала), либо по ансамблю (для случайных сигналов). Это равенство выполняется, в частности, при $A(t) = \text{const}$, т. е. для сигналов с чисто угловой модуляцией, и при $\Omega(t) = \text{const}$, т. е. для сигналов с чисто амплитудной модуляцией. Кроме того, оно выполняется при некоррелированных значениях A^2 и Ω в одинаковые моменты, когда $<A^2(t)\,\Omega(t)> = <A^2(t)><\Omega(t)>$. Это имеет место для гауссовских сигналов с симметричным спектром, не содержащих регулярной составляющей.

Если бы не существовало теоремы 2, то мгновенная частота могла бы сколь угодно отклоняться от своего среднего значения Ω_I независимо от изменения огибающей, лишь бы эти отклонения распределялись в известном смысле симметрично

относительно Ω_I. Вторая теорема ограничивает эти возможности, но не абсолютно. Отклонения могут быть действительно любыми и даже несимметричными, как в приведенном ранее примере суммы двух гармонических составляющих с близкими частотами и амплитудами. Но большие отклонения мгновенной частоты обязательно сопровождаются значительными уменьшениями огибающей, обеспечивающими выполнение неравенства $\Omega^2_{II} \leqslant |\omega^2_{II}$, вытекающего из (6.18). Поэтому же большие отклонения частоты не могут быть длительными — вспомним предположение D.

Из (6.19) можно сделать важное качественное заключение. Средний квадрат ширины спектра произвольного сигнала состоит из двух частей, одна из которых определяется угловой модуляцией и равна среднему квадрату девиации, а вторая — амплитудной модуляцией. Ширина спектра $\delta\omega$ может равняться нулю только в отсутствие как угловой, так и амплитудной модуляции, т. е. только для гармонического сигнала. Интересно отметить, что для нормального случайного процесса с симметричным спектром обе части $(\delta\omega)^2$, а именно $(\delta\Omega)^2$ и $(\delta_A\omega)^2 = \lim\limits_{T\to\infty} \left[\int\limits_{-T}^{T} A'^2(t)\,dt \Big/ \int\limits_{-T}^{T} A^2(t)\,dt\right]$, одинаковы, каждая из них равна $0{,}5(\delta\omega)^2$, так что для нормального шума ширина спектра в одинаковой мере определяется изменениями его огибающей и фазы.

Минимальную ширину спектра при заданной огибающей имеет сигнал с постоянной мгновенной частотой, для которого $\Omega_I = \Omega_{II} = \Omega$ и $\delta\Omega = 0$. Аналогично при заданном законе угловой модуляции минимальную ширину спектра имеет сигнал с постоянной амплитудой, когда $A'(t) = 0$.

6.5. Разочарование

В течение десяти лет автор очень гордился полученными ф-лами (6.17) — (6.19). Они оказались полезными в некоторых случаях для решения практических задач и к тому же были достаточно красивы. А получить красивый математический результат — это большое удовольствие особенно для не математика.

И лишь после того, как эта глава была написана, автор решил еще раз перелистать работу [35], которую он читал в 1950 г., и — о ужас! Он нашел в ней эти самые формулы!

Правда, они получены там только для периодических сигналов. Но теперь уже единственное, на что может автор претендовать — это на обобщение известного ранее результата и на его популяризацию.

Ничего не поделаешь. Лишний раз убеждаешься, что новое — это хорошо забытое старое. А также, что размышляя над проблемой, не вредно почитывать литературу.

6.6. Кое-что о линейных цепях

Действительная и мнимая части передаточной функции $K(i\omega)$ физически реализуемой линейной цепи, как известно, связаны преобразованием Гильберта. Поэтому все результаты, полученные для гильбертовского сигнала, можно применить и к линейным цепям, и наоборот. При этом нужно только иметь в виду, что в выражении для гильбертовского сигнала аргументом является время, а для передаточной функции — частота. Соответственно преобразованием Фурье для сигнала является его спектральная плотность, аргументом которой слу-

жит частота, а для передаточной функции преобразованием Фурье является импульсная характеристика $h(t)$ с временем в качестве аргумента. Обратим внимание на то, что для физически реализуемой цепи $h(t)=0$ при $t<0$ так же, как для аналитического сигнала $S(\mathrm{i}\omega)=0$ при $\omega<0$.

Продолжая эту аналогию, можно составить таблицу соответствия параметров, характеризующих цепи и сигналы[1]:

Сигнал	*Физически реализуемая цепь*
Время t	Частота ω
Частота спектра ω	Время t
Гильбертовский сигнал $\hat{u}(t)$	Передаточная функция $K(\mathrm{i}\omega)$
Спектральная плотность $S(\mathrm{i}\omega)$	Импульсная характеристика $h(t)$
Огибающая $A(t)$	Амплитудно-частотная характеристика (АЧХ) $\lvert K(\mathrm{i}\omega)\rvert = C(\omega)$
Полная фаза $\Phi(t)$	Фазо-частотная характеристика (ФЧХ) $\varphi(\omega) = \arg K(\mathrm{i}\omega)$
Мгновенная частота $\Omega(t)=\dfrac{d\Phi}{dt}$	Групповая задержка $\tau(\omega)=\dfrac{d\varphi}{d\omega}$
Средняя частота спектра $\omega_I = \dfrac{\int_0^\infty \omega \lvert S(\mathrm{i}\omega)\rvert^2 d\omega}{\int_0^\infty \lvert S(\mathrm{i}\omega)\rvert^2 d\omega}$	Среднее запаздывание импульсного отклика $t_I = \dfrac{\int_0^\infty t h^2(t)\, dt}{\int_0^\infty h^2(t)\, dt}$

[1] Для простоты ограничимся сигналами из L_2, для которых существует спектральная плотность $S(\mathrm{i}\omega)$. Это позволяет также обойтись без предельного перехода в ф-лах (6.14), (6.15) и др.

Средний квадрат частоты спектра

$$\omega_{II}^2 = \frac{\int_0^\infty \omega^2 \,|\, S(i\omega) \,|^2 \, d\omega}{\int_0^\infty |\, S(i\omega) \,|^2 \, d\omega}$$

Средний квадрат ширины спектра

$$(\delta\omega)^2 = \omega^2{}_{II} - \omega^2{}_I$$

Средняя мгновенная частота

$$\Omega_I = \frac{\int_{-\infty}^\infty A^2(t)\,\Omega(t)\,dt}{\int_{-\infty}^\infty A^2(t)\,dt}$$

Средний квадрат мгновенной частоты

$$\Omega_{II}^2 = \frac{\int_{-\infty}^\infty A^2(t)\,\Omega^2(t)\,dt}{\int_{-\infty}^\infty A^2(t)\,dt}$$

Средний квадрат девиации

$$(\delta\Omega)^2 = \Omega^2{}_{II} - \Omega^2{}_I$$

Средний квадрат запаздывания импульсного отклика

$$t_{II}^2 = \frac{\int_0^\infty t^2 h^2(t)\,dt}{\int_0^\infty h^2(t)\,dt}$$

Средний квадрат длительности импульсного отклика

$$(\delta t)^2 = t^2{}_{II} - t^2{}_I$$

Средняя задержка цепи

$$\tau_I = \frac{\int_0^\infty \tau(\omega)\,C^2(\omega)\,d\omega}{\int_0^\infty C^2(\omega)\,d\omega}$$

Средний квадрат задержки цепи

$$\tau_{II}^2 = \frac{\int_0^\infty \tau^2(\omega)\,C^2(\omega)\,d\omega}{\int_0^\infty C^2(\omega)\,d\omega}$$

Средний квадрат дисперсии[1]

$$(\delta\tau)^2 = \tau^2{}_{II} - \tau^2{}_I$$

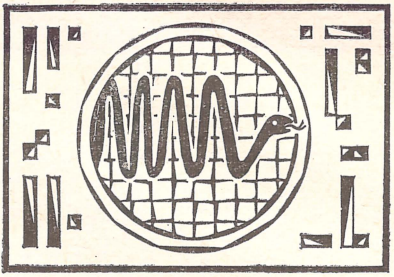
---

[1] Здесь термин «дисперсия» используется в «оптическом» смысле как разброс скорости прохождения волн с различными частотами.

По аналогии с теоремами, полученными для сигналов, можно записать следующие равенства:

$$\left.\begin{aligned} t_I &= \tau_I, \\ t_{II}^2 &= \tau_{II}^2 + \int_0^\infty C'^2(\omega)\,d\omega \Big/ \int_0^\infty C^2(\omega)\,d\omega, \\ (\delta t)^2 &= (\delta\tau)^2 + \int_0^\infty C'^2(\omega)\,d\omega \Big/ \int_0^\infty C^2(\omega)\,d\omega. \end{aligned}\right\} \qquad (6.21)$$

Непосредственный вывод этих формул в других обозначениях и с другой терминологией дан в [46], с. 510—513. Там же эти соотношения используются для доказательства того, что при заданной АЧХ $C(\omega)$ — наименьший средний квадрат длительности импульсного отклика — имеет цепь с линейной ФЧХ, для которой $\tau(\omega) = \text{const}$ и, следовательно, дисперсия $\delta\tau = 0$. Этот результат полезен для корректирования характеристик канала с целью уменьшения межсимвольной интерференции.



7

ВОКРУГ ЧАСТОТНОЙ МОДУЛЯЦИИ

7.1. Применяется ли частотная модуляция на практике?

Прежде всего, что такое частотная модуляция? Просматривая разные старые работы, с удивлением замечаешь, что подход к рассмотрению любой модуляции претерпел за последние 20—30 лет глубокие изменения. В настоящее время, говоря о том или ином виде модуляции, мы прежде всего рассматриваем требуемый конечный результат. Так, например, при фазовой модуляции (ФМ) несущего колебания $A\cos\omega_0 t$ первичным сигналом $x(t)$ требуется синтезировать сигнал

$$s(t) = A\cos[\omega_0 t + \mu x(t)]. \qquad (7.1)$$

При частотной же модуляции (ЧМ) синтезируется сигнал

$$s(t) = A\cos[\omega_0 t + \mu\psi(t)], \qquad (7.2)$$

где $\frac{d\psi}{dt} = x(t)$, и т. п.

После того как сформулирована цель, можно ставить вопрос о реализационной схеме, позволяющей осуществить такой синтез на практике.

Не так давно подход был другим. Все начиналось со схемы. В частности, ЧМ определялась как операция воздействия сигнала $x(t)$ на конденсатор или катушку контура автогенератора, в результате чего емкость или индуктивность изменяется по закону $C(t)=C_0+\Delta Cx(t)$, а затем доказывалось, что получаемое колебание при определенных условиях принимает в первом приближении форму (7.2). При таком старом подходе модуляция путем изменения резонансной частоты контура автогенератора всегда остается частотной независимо от закона изменения резонансной частоты[1].

Будем придерживаться современного определения ЧМ и других видов угловой модуляции, в соответствии с которым сигнал $A\cos\Phi(t)$ считается модулированным по частоте сигналом $x(t)$ только в том случае, если $\frac{d\Phi}{dt}=\omega_0+\mu x(t)$, и несколько уточним поставленный в заголовке вопрос — находит ли в настоящее время широкое применение частотная модуляция для передачи речи и музыки.

Правильный ответ звучит неожиданно: в настоящее время частотная модуляция для передачи звуковых сообщений, в частности для радиовещания, почти никогда не применяется.

Как же так?! — воскликнет читатель. — У меня есть приемник 1-го класса с УКВ диапазоном, специально предназначенным для приема высококачественных ЧМ радиовещательных программ. В моем телевизоре прием звукового сопровождения ведется с помощью частотного детектора, значит, передается оно путем частотной модуляции. Да и в описании телевизора об этом говорится.

---

[1] Этот принцип классификации рекомендован, например, в [35].

Все это верно, если частотной модуляцией называть всякий вид модуляции, в котором мгновенная частота изменяется и несет информацию о передаваемом сообщении. Но сейчас для этого применяется термин «угловая модуляция», частотной же называется только та ее разновидность, при которой отклонения мгновенной частоты пропорциональны первичному (низкочастотному) сигналу $x(t)$. Этому соответствует ур-ние (7.2), тогда как (7.1) определяет фазовую модуляцию.

В современных передатчиках с угловой модуляцией первичный сигнал воздействует на контур задающего генератора, изменяя генерируемую частоту. Однако первичный (модулирующий) сигнал подается на управляющий элемент не непосредственно, а через корректирующую цепь, выполняющую так называемую операцию предыскажения. Сущность ее, как обычно пишут, заключается в подъеме усиления на верхних частотах спектра первичного сигнала. Принятый сигнал подвергается частотному детектированию, после чего поступает на восстанавливающую цепь, которая снижает усиление на верхних частотах, компенсируя тем самым предыскажения в передатчике. При пояснении этой последовательности обработки сигнала обычно говорят следующее.

На выходе частотного детектора шум, как известно, имеет параболическую спектральную плотность мощности, так что верхние частоты спектра модулирующего сигнала сильнее поражены шумом, чем нижние. С другой стороны, верхние частоты речевого спектра менее мощны, но более информативны, чем нижние. Поэтому их важно передать как можно вернее. Для этого осуществляют предыскажения, поднимая мощность верхних частот и повышая тем самым отношение спектральных плот-

ностей сигнал/шум на верхних частотах, конечно, за счет соответствующего снижения отношения на нижних частотах. В восстанавливающей цепи полученное отношение сохраняется, а исходная форма спектра восстанавливается.

Все как будто бы ясно, если не считать не вполне уместной ссылки на спектр речи и информативность верхних частот. Впрочем интересно, сохраняется ли при этих операциях общее отношение мощностей сигнала и помехи на выходе приемника. Ведь на верхних частотах в результате предыскажения оно повысилось, а на нижних — понизилось. Сохраняется ли общий баланс? И тут-то оказывается, что в системе с предыскажением и выравниванием суммарное отношение мощностей сигнал/помехи втрое хуже (!), чем при отсутствии предыскажений[1]. Почему же это выгодно?

Все становится значительно проще и яснее, если подойти к этому вопросу с другой стороны. Ведь обычное предыскажение (с повышением уровня на 6 дБ на октаву) есть не что иное, как дифференцирование сигнала. Оказывается, на частотный модулятор подается не сигнал $x(t)$, а напряжение, пропорциональное его производной dx/dt. Поэтому мгновенная частота модулированного вторичного сигнала пропорциональна dx/dt, а значит, мгновенная начальная фаза пропорциональна $x(t)$. Следовательно, это модуляция фазовая, соответствующая ф-ле (7.1). И детектируется принятый сигнал в конечном счете по фазе — сначала частотный детектор выделит напряжение, пропорциональное мгновенной частоте, т. е. dx/dt, а затем оно интегрируется в цепи восстановителя и в ре-

[1] См. сравнение выигрышей систем ЧМ и ФМ, например, в [15, 16, 19] и др.

зультате получается напряжение, пропорциональное мгновенной фазе.

В более общем случае первичный сигнал $x(t)$ поступает на предыскажающий фильтр, имеющий некоторую заданную передаточную функцию $K(i\omega)$, не обязательно равную или близкую к передаточной функции идеального дифференциатора $ia\omega \operatorname{sgn} \omega$. Видоизмененный этим фильтром первичный сигнал $y(t)$ модулирует вторичный сигнал по частоте. При этом ни мгновенная фаза, ни мгновенная частота не пропорциональны $x(t)$, следовательно, получившаяся угловая модуляция не является ни фазовой, ни частотной.

В большинстве случаев в радиовещании и в многоканальной радиорелейной и спутниковой связи с «частотной» модуляцией[1] используют предыскажение, близкое к дифференцированию, так что получаемая угловая модуляция близка к фазовой. Причина этого в многоканальных системах ясна — только при фазовой модуляции спектр нормального шума на выходе детектора равномерен в полосе частот модулирующего группового сигнала и поэтому на всех поднесущих обеспечивается одинаковая помехоустойчивость. В случае радиовещательного сигнала важно то обстоятельство, что для высокого качества художественного восприятия верхние частоты речевого или музыкального сигнала должны превышать уровень шума не в меньшей степени, чем нижние.

Итак, на вопрос, поставленный в заголовке раздела, можно ответить отрицательно — в чистом виде частотная модуляция практически не применя-

---

[1] Имеется в виду частотная модуляция несущей групповым сигналом, который может быть получен сложением любым способом промодулированных поднесущих.

ется, а применяемая угловая модуляция в большинстве случаев значительно ближе к фазовой, чем к частотной.

7.2. Ширина спектра при угловой модуляции

Теоретически ширина спектра при угловой модуляции бесконечна, даже если первичный (модулирующий) сигнал имеет финитный спектр. Но инженера интересует та полоса частот, в которой расположена существенная часть спектра и которую нужно по возможности передать без заметных искажений, чтобы затем после детектирования удовлетворительно восстановить первичный сигнал.

Этому вопросу уделяется внимание в учебной и справочной литературе. Очень часто при этом пользуются формулой Манаева [29], согласно которой максимальный порядок боковых полос в существенной части спектра определяется приблизительно, как $n=1+m+\sqrt{m}$, где m — индекс частотной модуляции.

Эта формула получена из предположения, что можно пренебрегать боковыми частотами, амплитуды которых в 100 и более раз меньше немодулированной несущей.

Против этой формулы иногда возражают исходя из того, что амплитуды отбрасываемых составляющих спектра следует сравнивать не с амплитудой немодулированной несущей, а с амплитудой максимальной сохраняемой составляющей. В [25] приведены формулы типа $2\Delta\omega^*=A\Delta\omega+B\Omega$, где $2\Delta\omega^*$ — ширина существенной части спектра; $\Delta\omega$ — девиация частоты; Ω — модулирующая частота; A и B — коэффициенты, зависящие от отношения максимальной амплитуды отбрасываемых

составляющих спектра к максимальной амплитуде сохраняемых.

Все эти формулы обладают существенным недостатком как с инженерной, так и с методической точек зрения — они получены путем исследования спектра ЧМ сигнала при модуляции *гармонической функцией* с частотой Ω. Это затрудняет использование формул в реальных условиях, когда первичный сигнал является сложным, и поэтому ни частота Ω, ни девиация $\Delta\omega$ строго не определены. В некоторых учебниках рекомендуют при этом вести расчет по верхней частоте $\Omega_{макс}$ первичного сигнала, что, однако, ничем не обосновано, поскольку при настоящей ЧМ индекс модуляции для нижних частот значительно выше, чем для верхних.

С методической точки зрения эти формулы опасны тем, что они приучают молодого инженера «мыслить синусоидой». В данном случае это не вызывает грубых ошибок, однако в дальнейшем будут приведены примеры, когда такое «синусоидальное мышление» приводит к нелепым результатам.

Попробуем подойти к вопросу о ширине спектра сигнала с угловой модуляцией, не ограничиваясь гармоническим первичным сигналом. При этом необходимо задаться определением меры ширины спектра, который, как уже отмечалось, по существу, бесконечен.

Для теоретических исследований наиболее удобной мерой является средняя квадратичная ширина спектра, определенная ф-лой (6.11). Если известен первичный сигнал и параметры модуляции, то средняя квадратичная ширина спектра получается из ф-лы (6.19) с учетом того, что для угловой модуляции $A'(t)=0$:

$$(\delta\omega)^2 = (\delta\Omega)^2, \qquad (7.3)$$

где для случайного первичного сигнала средний квадрат девиации $(\delta\Omega)^2$ определяется его вероятностными характеристиками и глубиной угловой модуляции.

Заметим, что, так как при угловой модуляции $A(t)=\text{const}$, ф-ла (6.16) упрощается и средний квадрат девиации определяется как $(\delta\Omega)^2=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}[\Omega(t)-\Omega_I]^2dt$ или для случайных сигналов $(\delta\Omega)^2=\langle[\Omega(t)-\Omega_I]^2\rangle$.

Полученный результат очень прост, однако он не удовлетворяет разработчика, которому нужно знать расчетную полосу пропускания цепи для прохождения ЧМ сигнала с допустимыми искажениями. Эта полоса обычно шире среднеквадратичной ширины спектра, так как за пределы последней попадают многие спектральные составляющие, отбрасывая которые можно заметно исказить ход мгновенной частоты. К сожалению, дать универсальные формулы для расчета такой полосы по средней квадратичной девиации принципиально невозможно, так как эта полоса зависит и от тонкой структуры первичного сигнала. Поэтому попробуем найти хотя бы оценки, позволяющие определить границы необходимой полосы пропускания.

Будем считать, что существенная часть спектра расположена в полосе частот, за пределами которой содержится лишь некоторая заданная небольшая доля ε энергии или мощности сигнала[1]).

Обозначим минимальную полосу частот, обладающую этим свойством $\Delta\omega_\varepsilon$. Пусть известно, что средний взвешенный квадрат девиации сигнала равен $(\delta\Omega)^2$ и больше ничего о модулирующем сигнале неизвестно. Найдем максимальное возможное значение $\Delta\omega_\varepsilon$.

Эту задачу удобнее решать «наоборот». Зададимся некоторым значением $\Delta\omega_\varepsilon$ и отыщем минимальное значение сред-

---

[1]) Обращаем внимание читателя на то, что такое определение коренным образом отличается от тех, где за пределами существенной части спектра допускается существование составляющих с амплитудами, не превышающими определенной величины, что имеет смысл только для дискретного спектра.

него квадрата девиации $(\delta\Omega)^2$, которое согласно (7.3) равно среднему квадрату ширины спектра $(\delta\omega)^2$. По определению

$$(\delta\omega)^2 = 2\int_0^{\infty} (\omega - \omega_I)^2 g(\omega)\, d\omega, \tag{7.4}$$

где $g(\omega)$ — нормированная энергетическая спектральная плотность. Требуется найти $g(\omega)$, минимизирующее $(\delta\omega)^2$ при условии

$$2\int_{\omega_I - \frac{\Delta\omega_\varepsilon}{2}}^{\omega_I + \frac{\Delta\omega_\varepsilon}{2}} g(\omega)\, d\omega = 1 - \varepsilon. \tag{7.5}$$

Здесь предполагается, что полоса частот, содержащая существенную часть спектра, расположена симметрично относительно средней частоты ω_I. Перепишем (7.4) так:

$$(\delta\omega)^2 = 2\int_0^{\omega_I - \frac{\Delta\omega_\varepsilon}{2}} (\omega - \omega_I)^2 g(\omega)\, d\omega + 2\int_{\omega_I - \frac{\Delta\omega_\varepsilon}{2}}^{\omega_I + \frac{\Delta\omega_\varepsilon}{2}} (\omega - \omega_I)^2 g(\omega)\, d\omega + 2\int_{\omega_I + \frac{\Delta\omega_\varepsilon}{2}}^{\infty} (\omega - \omega_I)^2 g(\omega)\, d\omega \tag{7.6}$$

и будем минимизировать полученную сумму при условии (7.5). Очевидно, что средний интеграл принимает минимальное значение, равное нулю, когда спектральная плотность $g(\omega)$ в пределах полосы $\Delta\omega_\varepsilon$ целиком сосредоточена на сред-

ней частоте ω_I. Крайние интегралы в сумме дадут наименьшее значение, когда остающаяся часть $g(\omega)$ сосредоточена на частотах $\omega_I \pm \Delta\omega_\varepsilon/2$. Итак, нормированная спектральная плотность, минимизирующая (8.4) при условии (8.5), имеет вид

$$g(\omega) = 0{,}5\,[\alpha\varepsilon\delta(|\omega| - \omega_I + \Delta\omega_\varepsilon/2) + (1-\varepsilon)\,\delta(|\omega| - \omega_I) + (1-\alpha)\,\varepsilon\delta(|\omega| - \omega_I - \Delta\omega_\varepsilon/2)],$$

где α — любое действительное число, такое, что $0 \leqslant \alpha \leqslant 1$. При этом

$$(\delta\omega)^2 = (\delta\Omega)^2 = \varepsilon\,(\Delta\omega_\varepsilon/2)^2$$

(рис. 7.1).

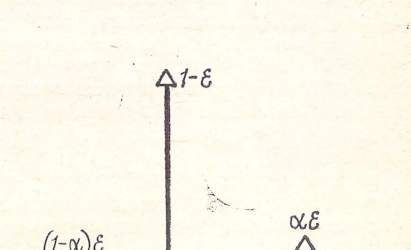


Рис. 7.1. Спектральная плотность мощности ЧМ сигнала с максимальным отношением ширины спектра к средней квадратичной девиации

Таким образом, ширина полосы $\Delta\omega_\varepsilon$ при фиксированной средней квадратичной девиации $\delta\Omega$ ограничена сверху величиной

$$\Delta\omega_\varepsilon < 2\delta\Omega/\sqrt{\varepsilon}. \qquad (7.7)$$

Обычно при расчетах удобнее пользоваться не средней квадратичной девиацией $\delta\Omega$, а ее пиковым значением $\delta\Omega_{макс} = \delta\Omega\Pi$, где Π — пикфактор первичного сигнала при ЧМ или его производной при ФМ. В общем случае угловой модуляции Π — пикфактор закона изменения мгновенной частоты. При этом

$$\Delta\omega_\varepsilon < 2\delta\Omega_{макс}/(\Pi\sqrt{\varepsilon}). \qquad (7.7а)$$

Для речевого сигнала $\Pi \approx 3 \div 4$, а при передаче музыки $\Pi \approx 5 \div 7$. Если задаваться значением $\varepsilon = 0{,}1$, то правая часть неравенства (8.7а) принимает значение порядка $(1 \div 2)\, \delta\Omega_{\text{макс}}$, если же определять полосу пропускания, исходя из более жесткого требования $\varepsilon = 0{,}01$, то она возрастает до $(3 \div 6)\, \delta\Omega_{\text{макс}}$.

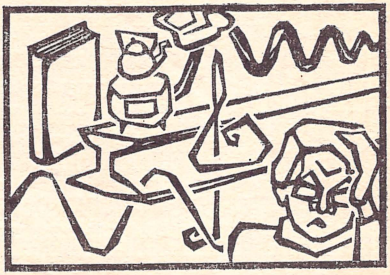
Граница (7.7) практически достигается при малых индексах модуляции. Чаще возникает задача оценки ширины спектра сигнала при больших индексах угловой модуляции, т. е. при медленном (по сравнению с величиной девиации) изменении мгновенной частоты. В этих условиях допустим квазистационарный подход. Пользуясь методом стационарной фазы (см., например, [5]), можно показать, что энергетическая спектральная плотность ЧМ сигнала на частоте ω при большом среднем индексе модуляции примерно пропорциональна времени нахождения мгновенной частоты $\Omega(t)$ в интервале $(\omega, \omega + d\omega)$. В частности, при модуляции случайным первичным сигналом, если индекс модуляции велик, форма энергетической спектральной плотности приблизительно совпадает с формой одномерной плотности вероятности модулирующего сигнала.

Исходя из этих соображений, ширина спектра сигнала с угловой модуляцией может быть определена приближенно как удвоенная максимальная девиация:

$$\Delta\omega_{\varepsilon} \approx 2\delta\Omega_{\text{макс}}. \qquad (7.8)$$

Это равенство справедливо при значениях ε, тем меньших, чем больше средний индекс модуляции. Робинсон, запатентовавший узкополосную частотную модуляцию (см. гл. 2), ошибался «только» в том, что применил эту формулу при очень малом индексе.

8



НА ОДНОЙ БОКОВОЙ

8.1. Рекорды однополосной модуляции

— Какой из известных видов модуляции можно считать самым простым? — спрашивает профессор на экзамене.

— Однополосную модуляцию,— отвечает студент. Она представляет собой просто перенос спектра первичного сигнала из области низких частот в область высоких частот. К этому еще иногда добавляется инверсия спектра, если передается нижняя боковая полоса. Модуляция и детектирование здесь сводятся, по существу, к преобразованию частоты. Однополосная модуляция линейна, т. е. к ней применим принцип суперпозиции. При однополосной модуляции сохраняется ширина спектра сигнала и не изменяется отношение сигнал/помеха. Все это свидетельствует о крайней простоте однополосной модуляции.

— Вы совершенно правы, — говорит профессор. — Могу поставить вам «отлично». Кто следующий?

— Я, — говорит подошедший другой студент.

— Вот вам вопрос. Какой из известных вам видов модуляции можно считать самым сложным?

— Пожалуй, однополосную модуляцию. При однополосной модуляции изменяются и огибающая сигнала, и его мгновенная фаза, и его мгновенная частота, так что в ней объединяются и амплитудная и угловая модуляции. Но ни огибающая, ни мгновенная фаза, ни мгновенная частота, ни какой-либо другой параметр вторичного сигнала не повторяет первичный сигнал, как это имеет место почти во всех остальных видах модуляции. Сам процесс однополосной модуляции чрезвычайно сложен. Обычно сначала на низкой поднесущей частоте осуществляется балансная модуляция и фильтром выделяется нужная боковая полоса, затем этот процесс повторяется на более высокой поднесущей и т. д., пока не будет достигнуто необходимое положение спектра на оси частот. Существуют и другие способы, но все они, кажется, еще более сложны. Для восстановления формы первичного сигнала приходится применять также сложную процедуру демодуляции — восстановить подавленную несущую частоту, к тому же в определенной фазе, и детектировать сигнал вместе с ней. При этом не легко обеспечить неискаженное детектирование. Для восстановления несущей обычно приходится применять пилот-сигнал. Одним словом, трудно перечислить все сложности, связанные с однополосной модуляцией.

— Вы совершенно правы, — говорит профессор. — Вам также ставлю «отлично».

— Но позвольте, — вмешался слушавший все это третий студент. Два диаметрально противоположных ответа вы оценили одинаково высоко. Какой же из них верен на самом деле?

— Оба верны. Здесь случай, когда явления че-

ресчур многогранны и к ним нельзя подходить с позиций формальной логики и «исключенного третьего». Такую же ситуацию еще в древности подметил Эзоп, когда ему поручили купить на рынке самое лучшее, а затем самое худшее. Как известно, он в обоих случаях купил языки. Язык позволяет людям общаться, обмениваться информацией, без языка невозможны ни наука, ни искусство, ни даже развитое мышление. Но язык же породил ложь, клевету, брань, доносы...

Оставим, однако, нашего профессора и его студентов и поговорим серьезно об однополосной модуляции. Если бы первым изобретенным видом модуляции была однополосная модуляция (ОМ), то она, вероятно, всеми воспринималась бы как самая простая. Ее бы тогда и не называли однополосной, так как никому бы в голову не пришла мысль строить сигнал с двумя боковыми полосами. Модуляцию описывали бы просто как перенос спектра первичного сигнала вверх, а демодуляцию — как возвращение вниз.

Но в действительности первым видом модуляции явилась амплитудная (АМ). Когда ее спектр был проанализирован и инженеры привыкли к тому, что АМ сигнал содержит несущую частоту и две боковые полосы, некоторым из них пришла мысль о том, что без ущерба для передаваемой информации и с немалой выгодой можно избавиться от несущей частоты, а также от одной из боковых полос. Так и была изобретена модуляция, которую вначале называли просто «передача на одной боковой полосе», сокращенно ОБП, а позднее, после того, как появились новые виды модуляции — фазовая (ФМ) и частотная (ЧМ), некоторые стали называть «амплитудная модуляция с одной боковой полосой». Ей противопоставлялись обычная

амплитудная модуляция, а также «амплитудная модуляция без несущей с двумя боковыми полосами» (ДБП).

Разумеется, сразу появились изобретатели, предложившие применять и фазовую однополосную модуляцию, и частотную однополосную модуляцию, осуществляемые, например, путем выделения фильтром части спектра, расположенной выше (или ниже) несущей частоты в обычном ФМ или ЧМ сигнале. Неясным оставался только вопрос, для чего это нужно. Спектр таких сигналов шире, а помехоустойчивость хуже, чем при обычной однополосной модуляции.

Хотя такой подход еще можно встретить в недавно изданных книгах, автор глубоко убежден, что он устарел. Для построения хорошей теории модуляции нельзя в основу классификации и определений ставить метод схемной реализации, как это делали примерно до 50—60-х годов, о чем уже ранее говорилось (см. стр. 105). Однополосную модуляцию (ОМ) не следует рассматривать как разновидность амплитудной. Как уже отмечалось, при ОМ изменяются и огибающая, и фаза, и мгновенная частота. Это особый вид модуляции, так же как и двухполосная балансная модуляция, имеющая свои особенности и довольно обширную область применения.

Мы видели, что ОМ можно рассматривать и как самую простую, и как самую сложную модуляцию. Ей принадлежат и некоторые другие рекорды, но среди них есть один печальный — ни об одном виде модуляции не было опубликовано столько взаимно противоречивых и ошибочных высказываний, как об ОМ. Некоторые из наиболее поучительных ошибок будут описаны далее. Но предварительно займемся определением ОМ.

8.2. Как записать однополосный сигнал?

Пусть задан первичный сигнал $x(t)$. Определить вид модуляции — это значит (с современной точки зрения) выразить через $x(t)$ вторичный сигнал $s(t)$, как это сделано, например, для ФМ в (7.1) или для ЧМ в (7.2). Соответствующая запись для ОМ может иметь следующий вид [см. ф-лу (3.27)]:

$$s_{\text{ОМ}}(t) = C\left[x(t)\cos(\omega_0 t+\varphi) \mp \hat{x}(t)\sin(\omega_0 t+\varphi)\right], \tag{8.1}$$

где C — произвольная постоянная; ω_0 — несущая частота; $\hat{x}(t)$ — преобразование Гильберта от $x(t)$; φ — начальная фаза подавленной несущей частоты.

В зависимости от выбора знака «—» или «+» в (8.1), получаются разновидности ОМ с верхней или с нижней боковой полосой.

Если в (8.1) ограничиться первым членом, то получим выражение для двухполосной модуляции без несущей (балансной модуляции)

$$s_{\text{ДБП}}(t) = Cx(t)\cos(\omega_0 t+\varphi). \tag{8.1а}$$

Более наглядная запись ОМ сигнала получается из (8.1), если представить первичный сигнал $x(t)$ в форме

$$x(t) = X(t)\cos\Phi(t), \tag{8.2}$$

где $X(t)$ — огибающая первичного сигнала, определенная по ф-ле (3.7). Тогда

$$s_{\text{ОМ}}(t) = CX(t)\cos[\omega_0 t \pm \Phi(t)], \tag{8.3}$$

откуда сразу видно, что ОМ сводится к сдвигу первичного сигнала вверх по оси частот на величину ω_0. Если в (8.3) выбран знак «—», то, помимо

сдвига, происходит также инверсия (замена знака мгновенной фазы), в результате чего передается не верхняя, а нижняя боковая полоса.

8.3. Как выглядит однополосный сигнал?

Речь пойдет, конечно, о том, как выглядит ОМ сигнал на экране осциллографа. Ответить на него очень просто. Если несущая частота ω_0 лежит значительно выше спектра первичного сигнала, что практически всегда выполняется, то в (8.3) $CX(t)$ является гильбертовской огибающей. Она и наблюдается при определенной скорости развертки в качестве граничной линии светлого участка экрана осциллографа. Очевидно, она совпадает с гильбертовской огибающей первичного сигнала (а не с самим первичным сигналом, как это было бы при АМ); это показано на рис. 8.1.

Предположим теперь, что в качестве первичного сигнала на однополосный модулятор подано напря-

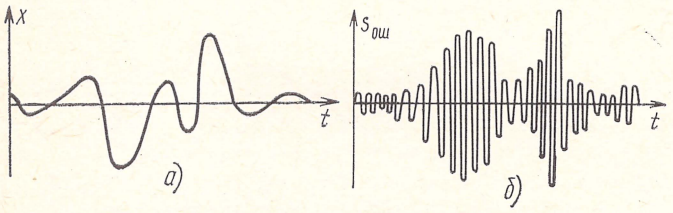


Рис. 8.1. Осциллограммы сигналов:
а) низкочастотного модулирующего; *б)* однополосного

жение от звукового генератора. Полагая, что генерируемое напряжение гармоническое, $x(t)=$ $=X\cos(\omega_M t+\Phi)$, и замечая, что его огибающая $X=\text{const}$, мы ожидаем получить на осциллографе

изображение ОМ сигнала, показанное на рис. 8.2, т. е. изображение гармонического сигнала.

Около 40 лет назад такой эксперимент был проделан в лаборатории, где работал автор этой книги. Это делалось не из любопытства — необходимо было проверить, правильно ли работает собранный макет однополосного модулятора. К наше-

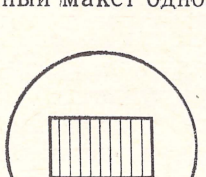


Рис. 8.2. Однополосный сигнал при модуляции гармоническим колебанием

Рис. 8.3. Фактически наблюдаемая осциллограмма при однополосной модуляции сигналом звукового генератора

му удивлению и огорчению полученное изображение выглядело так, как показано на рис. 8.3 — огибающая полученного сигнала была заметно промодулирована синусоидальной низкочастотной функцией. Модуляция имела глубину порядка нескольких процентов и легко наблюдалась при надлежащей синхронизации развертки.

Очень огорченные, мы решили, что модулятор работает плохо. Либо недостаточно подавлена несущая частота, либо остается заметный остаток от второй боковой частоты. Измерив частоту огибающей и обнаружив, что она совпадает с частотой ω_M первичного сигнала, мы отбросили предположение о плохом подавлении второй боковой частоты. Ведь в отсутствие несущей две боковые ча-

стоты дают биения с частотой $2\omega_M$. Поэтому все внимание было уделено подавлению несущей.

Сняв модулирующее напряжение, мы убедились, что остаток несущей частоты, «пролезающий» через балансный модулятор, ниже пикового уровня по меньшей мере на 60 дБ, так что глубина паразитной модуляции огибающей не должна была бы превышать 0,1%, в действительности же она была раз в 50 больше. Тогда возникло подозрение, что вследствие неодинаковых характеристик диодов, входящих в балансный модулятор, симметрия схемы, установленная в отсутствие модулирующего напряжения, нарушается при подаче последнего. Это предположение было проверено, но расчет, проведенный на основе измеренных реальных характеристик, показал, что возможная разбалансировка должна давать эффект раз в 15 меньше обнаруженного.

Лишь спустя несколько дней мы догадались о причине этого явления, которая оказалась весьма простой. Все дело в том, что первичный сигнал, снимаемый с звукового генератора, только номинально является гармоническим. Фактически он содержит, помимо основной гармонической составляющей с частотой ω_M, также высшие гармоники с частотами, кратными ω_M, причем их уровень даже в хороших современных лабораторных генераторах часто достигает 1—2%, а в те времена доходил до 5—6%. Поэтому и сформированный однополосный сигнал даже при идеальном подавлении несущей оказывается не чисто гармоническим и в его спектре, помимо частоты $\omega_0+\omega_M$, присутствуют $\omega_0+2\omega_M$, $\omega_0+3\omega_M$ и т. д. Биения между ними происходят по закону, близкому к гармоническому, с частотой ω_M.

В этом легко убедиться, ограничившись для

простоты только 2-й гармоникой и приняв амплитуду 1-й гармоники за единицу. Тогда

$$x(t) = \cos\omega_M t + \alpha\cos(2\omega_M t + \psi),$$

где α — величина порядка нескольких сотых; ψ — произвольный сдвиг фазы. Сопряженный сигнал выражается формулой

$$\hat{x}(t) = \sin\omega_M t + \alpha\sin(2\omega_M t + \psi)$$

и огибающая первичного сигнала

$$X(t) = \sqrt{x^2(t) + \hat{x}^2(t)} = \sqrt{1 + \alpha^2 + 2\alpha\cos(\omega_M t + \psi)} \approx$$

$$\approx 1 + \alpha\cos(\omega_M t + \psi).$$

Подставив это выражение в (8.3), видим, что огибающая однополосного сигнала промодулирована и глубина модуляции равна α.

Огибающую $X(t)$ первичного сигнала непосредственно на осциллографе наблюдать нельзя, так как этот сигнал не является узкополосным, а в этом случае «наглядность» огибающей отсутствует. Но при однополосной модуляции формируется узкополосный сигнал с той же огибающей и тут-то она и проявляется в явном, наглядном виде и иногда, как в описанном случае, вносит смятения в умы малоопытных исследователей.

8.4. «Формула Костаса»

Начиная с появления ОМ, в учебниках, журнальных статьях и монографиях усиленно дебатировался вопрос о том, какой выигрыш дает переход от АМ к ОМ. По этому вопросу было высказано много разноречивых мнений. В начале 60-х годов американский ученый Дж. Костас писал, что, просмот-

рев обширную журнальную литературу по однополосной модуляции, он обнаружил в каждой статье свою оценку энергетического выигрыша относительно амплитудной модуляции, от 2 до нескольких десятков и более. В результате анализа он установил, что выигрыш в децибелах, указываемый в каждой статье, составляет примерно $(3+N!)$ дБ, где N — число соавторов данной статьи.

Если эта шутка и не точна, она все же правильно отражает тот разнобой, который существовал в те годы. Помимо того, что разные авторы производили сравнение в различных условиях и по-разному определяли энергетический выигрыш, они также допускали немало различных ошибок.

Вот примеры некоторых рассуждений.

1. При обычной АМ, полагая мощность несущей частоты равной 1, имеем мощность пары боковых частот, равную $m^2/2$. При максимальном коэффициенте модуляции $m=1$ мощность боковых частот равна 1/2, т. е. составляет 1/3 от полной мощности сигнала. Остальные 2/3 мощности тратятся на несущую частоту, которая не несет информации и в этом смысле бесполезна. При переходе от АМ к ОМ вся мощность сигнала затрачивается на полезную боковую полосу. Следовательно, выигрыш равен 3.

2. При обычной АМ коэффициент модуляции m в среднем не превышает 0,45. Поэтому мощность боковых полос в среднем составляет 1/10 от мощности несущей или 1/11 от полной мощности сигнала. Следовательно, при переходе к ОМ полезная мощность возрастает в среднем в 11 раз.

3. В предыдущем рассуждении не учтено то, что в передатчике ограничена пиковая мощность, которая в моменты, когда $m=1$, достигает значения 3/2 от мощности несущей. Поэтому средняя мощность боковых частот при $m_{ср}=0{,}45$ составляет 1/10 от мощности несущей или 1/15 от пиковой мощности передатчика. Перейдя к ОМ, можно использовать всю пиковую мощность передатчика, т. е. получить выигрыш в 15 раз.

4. В предыдущем рассуждении не учитывалось, что однополосный сигнал имеет такой же пикфактор, как и модулирующий первичный сигнал. Пусть пикфактор примерно равен $П=3$. Следовательно, средняя мощность ОМ сигнала в $П^2=9$

раз меньше пиковой мощности передатчика, тогда как при АМ средняя мощность полезных боковых полос в 15 раз меньше пиковой мощности (см. предыдущее рассуждение). Таким образом, энергетический выигрыш составляет всего $15/9 \approx 1{,}67$ раза.

5. При АМ выигрыш в отношении сигнал/шум на выходе идеального приемника (по отношению к входу) равен $B = m^2/(1+m^2/2)$ или при $m=1$ равен 2/3. Далее цитирую буквально, по монографии [37]:

«При расчете выигрыша отношение полос частот[1] было взято равным 2. Фактически это отношение несколько больше, вследствие чего значение B ближе к единице. В практике часто принимают значение B при АМ равным единице (?!).

«Анализ обычных методов приема при АМ и линейном детектировании в случае малых шумов показывает, что по помехоустойчивости этот способ не отличается от идеального. Анализ обычных методов приема при синхронном детектировании, при квадратурной модуляции, при работе без несущей с двумя боковыми полосами и при одной боковой полосе также показывает, что и в этих случаях достигается потенциальная помехоустойчивость. Отсюда следует, что для этих систем коэффициент $B=1$».

Продолжая эту мысль, приходим к заключению, что никакого выигрыша ОМ относительно АМ не дает.

6. Будем сравнивать АМ и ОМ сигналы при одинаковой пиковой мощности. Это значит, что максимальные значения амплитуд в обоих случаях равны. Но максимум амплитуды АМ сигнала равен $U_н(1+m)$, где $U_н$ — амплитуда несущей. При $m=1$ половина максимальной амплитуды приходится на долю несущей частоты, а половина — на долю боковых. При переходе к ОМ вся максимальная амплитуда используется боковой частотой. Поэтому ее мощность оказывается в 4 раза больше, чем мощность боковых при АМ. Таков выигрыш в передатчике. Но к этому нужно прибавить еще выигрыш в приемнике. Последний обусловлен тем, что ОМ сигнал занимает вдвое более узкую полосу частот, чем АМ сигнал, и, следовательно, мощность шума при переходе к ОМ уменьшается в 2 раза. Таким образом, общий выигрыш равен 8.

7. К выигрышу, определенному предыдущим рассуждением, нужно добавить еще дополнительный выигрыш в случае коротковолновой радиосвязи. При связи отраженными волнами имеют место селективные замирания, которые в слу-

---

[1] Имеются в виду полосы на входе и выходе детектора.

чае АМ могут значительно исказить передачу, так как верхняя и нижняя боковые частоты могут получать различный сдвиг фазы относительно несущей. Это снижает средний уровень продетектированного сигнала и вносит дополнительные искажения, до некоторой степени эквивалентные увеличению шума. В то же время при ОМ селективные замирания могут только изменить соотношение амплитуд и фаз отдельных составляющих спектра восстановленного сигнала, что мало влияет на его разборчивость. Это преимущество ОМ можно оценить как дополнительный выигрыш в 2 раза, так что общий выигрыш оказывается равным 16.

Последняя цифра — выигрыш в 10 раз, или на 12 дБ, хотя и очень туманно обоснованная, почему-то получила наибольшее распространение и изредка появляется в печати даже сейчас, чаще всего без всяких обоснований, как якобы давно установленный факт.

Число примеров можно было бы еще умножить, но и этого достаточно. Все приведенные рассуждения страдают в основном одним недостатком — отсутствием четкой постановки задачи. Поскольку условия применения модуляции бывают разные, то и характеризовать выигрыш ОМ относительно АМ одним числом невозможно.

Современная теория связи понимает под энергетическим выигрышем некоторой системы A относительно системы B число, показывающее, во сколько раз нужно увеличить среднюю мощность сигнала в системе B, чтобы получить в обеих системах одинаковый результат на выходе. В идеальном случае, когда сигнал в канале не подвержен замираниям, несущая частота при приеме в точности известна, помехой является аддитивный белый гауссовский шум, а моделью сигнала служит стационарный случайный процесс, выигрыш ОМ относительно АМ, вычисленный на основании теории потенциальной помехоустойчивости при оптимальной обработке сигналов в обоих случаях, оказывается равным $(m^2+\Pi^2)/m^2$, где m — коэффициент

модуляции, П — пикфактор модулирующего сигнала. Наименьший выигрыш имеет место при $m=1$ и равен

$$B = 1 + П^2. \qquad (8.4)$$

Если $П \approx 3$, то $B \approx 10$. Такой же в точности выигрыш дает система ДБП (две боковые полосы без несущей) относительно АМ (см., например, [15, 19]).

Заметим, что в этих условиях никакой разницы в помехоустойчивости между системами с одной и с двумя боковыми полосами нет, хотя при ДБП спектр вдвое шире и поэтому мощность шума, неизбежно попадающего в приемник, вдвое больше, чем при ОМ. Это легко объяснить тем, что в процессе детектирования ДБП составляющие сигнала, расположенные симметрично относительно несущей, складываются когерентно, а сопутствующие им составляющие помехи — некогерентно. Это дает увеличение отношения сигнал/помеха в 2 раза, в точности компенсирующее добавление второй полосы помех. Кстати, это обстоятельство совершенно не учитывалось в «рассуждении 6», получившем широкое распространение, где сокращение полосы пропускания приемника вдвое при переходе к ОМ расценивалось как соответствующий выигрыш в помехоустойчивости.

Иначе обстоит дело, когда фаза сигнала в канале флуктуирует, тем более, когда флуктуирует мгновенная частота (например, вследствие нестабильности частоты передатчика или эффекта Допплера). В этом случае для приема сигнала ОМ с восстановлением формы передаваемого сообщения необходимо использование пилот-сигнала, содержащего информацию о частоте и фазе несущей. При обычной АМ несущая частота сама присутст-

вует в принимаемом сигнале, а при двухполосной модуляции без несущей она легко восстанавливается по боковым полосам [42]. Вследствие затраты лишней мощности на передачу пилот-сигнала, а также вследствие меньшей точности отслеживания несущей система ОМ проигрывает системе ДБП, а ее выигрыш относительно обычной АМ уменьшается. Степень этого уменьшения различна в зависимости от свойств первичного сигнала и от закона флуктуации фазы. Для первичных сигналов, моделируемых марковским случайным процессом, а также при описании флуктуаций фазы марковским процессом, результаты можно найти в [42], где дана также обширная библиография. Можно отметить, что по сравнению со случаем, когда фаза и частота не флуктуируют, величина выигрыша ОМ относительно АМ меняется незначительно.

В коротковолновой магистральной радиосвязи, когда вследствие перегруженности диапазона основной причиной нарушения связи являются взаимные помехи, а флуктуационный шум относительно мал, нужно совершенно иначе подходить к оценке преимуществ ОМ. В этом случае задачу можно сформулировать так: насколько изменится вероятность поражения сигнала мощной станционной помехой при переходе от АМ к ОМ. Совершенно очевидно, что эта вероятность уменьшится, но степень уменьшения зависит от многих факторов — от плотности помех, от распределения ширины спектра помехи и т. д.

При частотном уплотнении каналов с пренебрежимо малым уровнем помех ОМ позволяет увеличить кратность уплотнения вдвое по сравнению с АМ. Говорить об энергетическом выигрыше в этом случае не имеет смысла.

Если считать заданной не среднюю, а пиковую

мощность, то вычисленный выше выигрыш $1+\Pi^2$ нужно умножить на отношение квадратов пикфакторов АМ сигнала $\Pi_{АМ}$ и ОМ сигнала $\Pi_{ОМ}$, так что выигрыш ОМ по пиковой мощности относительно АМ составляет $B_{пик}=(1+\Pi^2)\Pi^2_{АМ}/\Pi^2_{ОМ}$. Напомним, что Π — пикфактор первичного сигнала.

Величину $\Pi_{АМ}$ легко выразить через Π. Из представления АМ сигнала

$$s_{АМ}(t) = A[1 + mx(t)]\cos\omega_0 t \qquad (8.5)$$

легко найти пиковое значение

$$s_{АМ}(t)_{макс} = A[1 + mx_{макс}],$$

или если первичный сигнал $x(t)$ нормирован так, что $x_{макс}=1$, то

$$s_{АМ}(t)_{макс} = A(1 + m). \qquad (8.6)$$

Средний квадрат АМ сигнала

$$s^2_{АМ}(t) = A^2\,\overline{[1 + m^2x^2(t) + 2mx(t)]\cos^2\omega_0 t} \approx$$

$$\approx A^2[1 + m^2\overline{x^2(t)} + 2m\,\overline{x(t)}]\;\overline{\cos^2\omega_0 t}.$$

Здесь мы воспользовались тем, что огибающая $A[1+mx(t)]$ меняется значительно медленнее высокочастотного заполнения $\cos\omega_0 t$, что позволяет усреднять $\cos^2\omega_0 t$, полагая значение огибающей за период несущей частоты постоянным.

Далее, обычно $\overline{x(t)}=0$, так как первичный сигнал не содержит постоянной составляющей, и $\overline{x^2(t)}=1/\Pi^2$, поскольку $x_{макс}=1$. Поэтому, учитывая, что $\overline{\cos^2\omega_0 t}=0{,}5$, получим:

$$s^2_{АМ}(t) = 0{,}5A^2(1 + m^2/\Pi^2), \qquad (8.7)$$

$$\Pi^2_{АМ} = (s_{АМ}(t)_{макс})^2/s^2_{АМ}(t) = 2(1 + m)^2/(1 + m^2/\Pi^2) =$$

$$= 2\Pi^2(1 + m)^2/(\Pi^2 + m^2). \qquad (8.8)$$

В наиболее интересном случае, когда $m=1$,

$$\Pi^2_{\text{АМ}} = 8\Pi^2/(1+\Pi^2). \qquad (8.9)$$

Что же касается однополосного сигнала, выражаемого ф-лой (8.3), то аналогичное вычисление дает

$$\Pi^2_{\text{ОМ}} = [X(t)_{\text{макс}}]^2/0{,}5X^2(t)], \qquad (8.10)$$

где $X(t)$ — огибающая первичного сигнала.

Если первичный сигнал (8.9) достаточно «узкополосен», так что при вычислении его среднего квадрата $\overline{x^2(t)} = \overline{X^2(t)\cos^2\Phi(t)}$ можно раздельно усреднить $X^2(t)$ и $\cos^2\Phi(t)$, причем $\overline{\cos^2\Phi(t)} = 0{,}5$, то легко видеть, что $\Pi = \Pi_{\text{ом}}$. Это приблизительно верно при телефонной модуляции. В этом случае

$$B_{\text{пик}} = (1+\Pi^2)\,8\Pi^2/(\Pi^2+1)\,\Pi^2 = 8, \qquad (8.11)$$

что совпадает с цифрой, полученной при интуитивной оценке примерно в тех же условиях. Однако при модуляции более широкополосным первичным сигналом обычно $\Pi_{\text{ом}} > \Pi$ и выигрыш по пиковой мощности оказывается меньше 8.

Предоставим читателю самостоятельно убедиться, что при двухполосной модуляции без несущей ф-ла (8.11) справедлива для выигрыша по пиковой мощности (относительно АМ), каков бы ни был первичный сигнал, лишь бы его спектр лежал существенно ниже несущей частоты.

8.5. Как детектировать однополосный сигнал?

Рассмотрим сначала другой вопрос — как вскипятить чайник. Точнее сформулируем задачу так: имеются пустой чайник, водопроводный кран, незажженная газовая плита и спички. Требуется по-

лучить чайник с кипящей водой. Решение этой задачи достаточно известно. Нужно, во-первых, наполнить чайник водой из крана, во-вторых, зажечь газ, в-третьих, поставить чайник на огонь и, наконец, подождать некоторое время, пока вода закипит.

Один математик, вполне усвоивший это решение, оказался в несколько иной ситуации — вода в чайнике уже была, а газ был зажжен. Как быть в таком случае? Очень просто, — сказал он. — Я выливаю воду из чайника и гашу газ, после чего задача сводится к предыдущей, решение которой нам уже известно.

Этот анекдот невольно приходит в голову, когда читаешь в некоторых книгах, что для приема ОМ сигнала (а также и для приема двухполосного сигнала без несущей) необходимо добавить к принятому сигналу несущую, превратить его тем самым в АМ сигнал, а затем уж задача сводится к известной — к детектированию АМ сигнала с помощью амплитудного детектора.

В случае добавления восстановленной несущей к двухполосному сигналу ее фаза должна точно совпадать с фазой «подавленной» несущей. Тогда верхние и нижние боковые частоты будут расположены симметрично относительно несущей, как показано на векторной диаграмме[1] рис. 8.4*а*. В противном случае, когда фаза несущей восстановлена неверно (рис. 8.4*б*), сумма принятого сигнала и восстановленной несущей оказывается промодулированной как по амплитуде, так и по фазе. Это видно из рис. 8.4*б*, где пунктиром показано перемещение конца результирующего вектора, который

[1] Для простоты здесь показан случай модуляции одним тоном. Предполагается, что плоскость чертежа вращается с угловой скоростью ω_0.

изменяет не только свою длину, но и направление. При этом огибающая, выделяемая амплитудным детектором, искажается, т. е. не повторяет форму

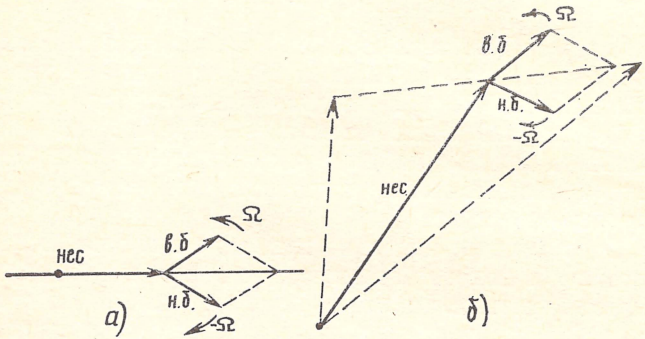


Рис. 8.4. Двухполосный сигнал с восстановленной несущей частотой при установке фазы:
а) точной; *б)* неточной

первичного сигнала, а ее величина уменьшается по сравнению со случаем, когда фаза несущей восстановлена точно.

На рис. 8.5 показана зависимость коэффициента гармоник $\kappa_г$ от неточности ψ восстановления фа-

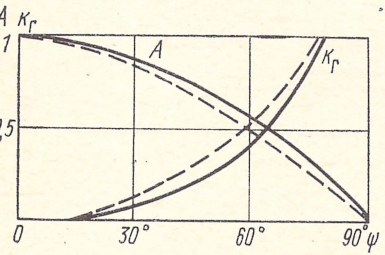


Рис. 8.5. Коэффициент гармоник $\kappa_г$ и амплитуда A продетектированного двухполосного сигнала с восстановленной несущей в зависимости от погрешности фазы несущей:
———— $E_н/E_б=20$;
— — — $E_н/E_б=2$

зы несущей при детектировании двухполосного сигнала амплитудным детектором. Там же показано отношение амплитуды полезного сигнала на выходе детектора к ее значению при $\psi=0$. Кривые построены для двух значений отношения амплитуды восстановленной несущей к амплитуде одной боковой частоты: $U_{\text{нес}}/U_{\text{б}}=2$ (что соответствует коэффициенту модуляции 1) и $U_{\text{нес}}/U_{\text{б}}=20$.

Увеличивая амплитуду восстановленной несущей, можно снизить нелинейные искажения до допустимой величины, если неточность фазы ψ не очень близка к 90°, однако при любой амплитуде несущей допустимая величина ψ не должна превышать 20—30° для того, чтобы амплитуда продетектированного сигнала не была существенно снижена. Здесь следует обратить внимание на то, что шум на выходе детектора практически не зависит от отклонения фазы несущей ψ, так что при больших ψ ухудшается отношение сигнал/шум на выходе детектора.

Значительно проще и благополучнее детектирование двухполосного сигнала без несущей осуществляется синхронным детектором. Для этого сигнал, выраженный ф-лой (8.1а), умножается на напряжение опорного гетеродина («восстановленную несущую») $a\cos(\omega_0 t+\psi)$, где $\psi-\varphi=\Psi$ — погрешность при восстановлении фазы несущей:

$$s(t)\,a\cos(\omega_0 t+\psi)=Cax(t)\cos(\omega_0 t+\varphi)\cos(\omega_0 t+\psi)=$$
$$=\frac{aC}{2}x(t)\cos[2\omega_0 t+\varphi+\psi]+\frac{aC}{2}x(t)\cos(\varphi-\psi). \qquad (8.12)$$

Первый член правой части представляет высокочастотную составляющую произведения, которая отфильтровывается. Второй же член является результатом детектирования и пропорционален пер-

вичному сигналу $x(t)$. Таким образом, при синхронном детектировании двухполосного сигнала никаких искажений не возникает, каковы бы ни были амплитуда a и начальная фаза ψ восстановленной несущей. Неточность фазы ψ вызывает только уменьшение уровня продетектированного сигнала, пропорциональное $\cos(\varphi-\psi)$. Если потребовать, чтобы напряжение продетектированного сигнала уменьшалось не более чем на 10%, необходимо восстановить фазу несущей с точностью $\pm 15°$, если же считать допустимым снижение напряжения на 30%, то требуемая точность фазы составит $\pm 45°$.

Преимущество синхронного детектирования перед детектированием огибающей суммы сигнала и восстановленной несущей очевидно. При обычной АМ, когда в принимаемом сигнале уже содержится несущая, детектор огибающей наиболее прост и поэтому практически только он и используется. При передаче без несущей последнюю все равно приходится «восстанавливать», т. е. синтезировать в приемнике. После этого синхронное детектирование выполняется столь же просто, как и детектирование огибающей.

Приблизительно так же решается задача детектирования ОМ сигнала. Сложение с восстановленной несущей и амплитудное детектирование полученной суммы «линейным» детектором, вообще говоря, приводят к нелинейным искажениям, для уменьшения которых приходится увеличивать амплитуду восстановленной несущей. Применение синхронного детектора избавляет от нелинейных искажений. Впрочем сохраняются линейные искажения, заключающиеся в сдвиге фаз всех составляющих продетектированного сигнала на величину, противоположную погрешности фазы восста-

новленной несущей. Действительно, используя запись ОМ сигнала (8.3), получим после синхронного детектирования

$$s_{\text{ОМ}}(t)\cos(\omega_0 t+\psi)=CX(t)\cos[\omega_0 t+\Phi(t)+\varphi]\times$$

$$\times a\cos(\omega_0 t+\psi)=\frac{aC}{2}X(t)\cos[2\omega_0 t+\Phi(t)+\varphi+\psi]+$$

$$+\frac{aC}{2}X(t)\cos[\Phi(t)+\varphi-\psi]. \qquad (8.13)$$

Полезным здесь является второй низкочастотный член. Он с точностью до постоянного множителя представляет первичный сигнал $x(t)$ с фазами, сдвинутыми на $\varphi-\psi$.

Как уже говорилось, ОМ представляет собой, в сущности, транспонирование спектра сигнала вверх, а синхронное детектирование — транспонирование вниз. Как при всяком транспонировании спектра (преобразовании частоты), здесь сохраняется огибающая сигнала, мгновенная же частота сдвигается на величину $\pm\omega_0$, где ω_0 — частота гетеродина. Начальная фаза при этом также смещается на величину, равную начальной фазе гетеродина. Это хорошо видно из выражений (8.1), (8.3) и (8.13). Здесь полезно еще раз напомнить, что огибающая первичного сигнала отличается от формы сигнала. Последняя поэтому может существенно изменяться при изменении начальной фазы.

На рис. 8.6 для примера показано изменение формы сравнительно простого периодического сигнала при сдвиге фаз его составляющих на 45 и 90°. Поэтому если требуется передать точную форму первичного сигнала, то восстановленная несущая при ОМ должна точно совпадать по фазе с подавленной несущей. В этом случае принимать двухполосный сигнал ДБП проще, так как там допу-

стимы отклонения фазы восстановленной несущей в пределах по крайней мере $\pm 15°$.

Однако дело коренным образом меняется, если первичный сигнал $x(t)$ является звуковым, т. е. рассчитан на восприятие слухом. Слуховой анализатор человека (и, по-видимому, всех животных) устроен так, что он не воспринимает непосредствен-

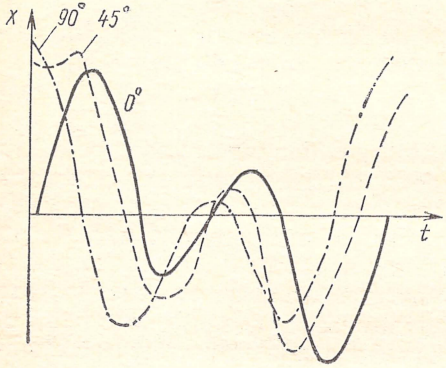


Рис. 8.6. Изменение формы низкочастотного сигнала при сдвиге фаз

но фазовых соотношений составляющих звука [1]. Поэтому какова бы ни была начальная фаза ψ опорного сигнала, звук воспринимается одинаково, лишь бы сохранился его амплитудный спектр. В этих условиях ОМ сигнал принимается проще, чем сигнал ДБП, так как допустимым является любое

[1] Впрочем, слух отличает звук, у которого фазовые соотношения гармонических составляющих остаются постоянными, от звука, у которого эти соотношения изменяются со временем. В первом случае воспринимается звук, идущий от одного источника, с обертонами. Во втором случае гармонические составляющие воспринимаются как исходящие от различных источников.

значение начальной фазы восстановленной несущей.

В ряде книг указывается, что при детектировании ОМ сигнала фаза восстановленной несущей может быть любой, но не оговаривается, что это справедливо только для звуковых сигналов. В частности, если бы при телевизионном вещании сигнал изображения формировался с помощью однополосной модуляции, то необходимо было бы в приемнике предусмотреть синхронный гетеродин, фаза которого каким-то образом устанавливалась по крайней мере с точностью до 1—2°. Это было бы трудно осуществить, и поэтому ОМ в телевидении не применяется. Для того же, чтобы сократить спектр излучаемого сигнала, в стандартном телевизионном сигнале одна из боковых полос частично отфильтровывается, но составляющая на несущей частоте сохраняется. После детектирования огибающей такого сигнала с частично подавленной боковой полосой первичный сигнал восстанавливается с существенными нелинейными искажениями, но при зрительном восприятии они практически незаметны. Такие искажения в основном вызывают изменения шкалы контрастности, но не отражаются на положении контуров изображения на экране и на общий характер переходов между частями изображения. Наоборот, линейные искажения в виде сдвигов фаз всех составляющих сигнала изменяют характер переходов, вызывают дробление точек, сглаживание резких контуров, иногда возникновение новых контуров и поэтому существенно нарушают зрительное восприятие.

При всяком преобразовании частоты существует возможность прохождения «зеркальной помехи», которую после преобразования невозможно подавить методами линейной частотной селекции, так

как ее спектр лежит в той же области частот, что и спектр полезного сигнала. Это же явление имеет место при обычном синхронном детектировании ОМ сигнала. При детектировании верхней боковой полосы зеркальная помеха расположена ниже несущей частоты и наоборот. Существует однако способ детектирования ОМ сигнала с подавлением зеркальной помехи непосредственно в детекторе. Для этого используются два перемножителя, в которых фазы опорного напряжения отличаются на 90°, широкополосный фазовращатель (преобразователь Гильберта)[1] и сумматор (рис. 8.7).

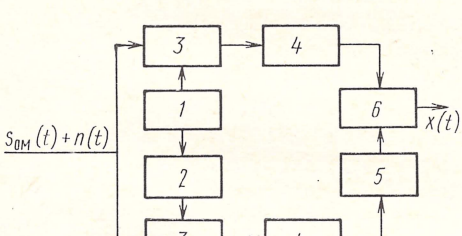


Рис. 8.7. Синхронный детектор однополосного сигнала с фазокомпенсационным подавлением зеркальной помехи:

1 — источник опорного напряжения; *2* — фазовращатель опорного напряжения; *3* — перемножитель; *4* — *ФНЧ*; *5* — широкополосный фазовращатель на 90° (преобразователь Гильберта); *6* — сумматор

Представим сигнал в виде (8.3), полагая для простоты $C=1$, а зеркальную помеху, лежащую по

[1] Хотя преобразование Гильберта является физически не реализуемой операцией, ее можно осуществить с любой заданной точностью, допустив достаточную задержку. Разумеется, эта задержка должна быть скомпенсирована такой же задержкой после другого перемножителя.

другую сторону от несущей, можно записать так:

$$n(t) = Y(t)\cos[\omega_0 t \mp \Psi(t)]. \qquad (8.14)$$

Примем для определенности, что передается верхняя боковая полоса, т. е. в ф-лах (8.3) и (8.14) выберем верхние знаки. На выходе 1-го перемножителя получим

$$[s_{ом}(t) + n(t)]\, a\cos(\omega_0 t + \psi) = 0{,}5aX(t)\cos[2\omega_0 t + \Phi(t) + \psi] + 0{,}5aY(t)\cos[2\omega_0 t - \Phi(t) + \psi] + 0{,}5aX(t)\cos[\Phi(t) - \psi] + 0{,}5aY(t)\cos[\Phi(t) + \psi],$$

а на выходе 2-го перемножителя

$$[s_{ом}(t) + n(t)]\, a\sin(\omega_0 t + \psi) = 0{,}5a\{X(t)\sin[2\omega_0 t + \Phi(t) + \psi] + Y(t)\sin[2\omega_0 t - \Psi(t) + \psi] - X(t)\sin[\Phi(t) - \psi] + Y(t)\sin[\Psi(t) + \psi]\}.$$

Выделив низкочастотные составляющие этих выходных сигналов и повернув фазы во втором из них на — 90°, получим

$$0{,}5X(t)\cos[\Phi(t) - \psi] + 0{,}5aY(t)\cos[\Psi(t) + \psi]$$

и

$$0{,}5aX(t)\cos[\Phi(t) - \psi] - 0{,}5aY(t)\cos[\Psi(t) + \psi]. \qquad (8.15)$$

При написании этих выражений учтено то, что косинус является четной функцией и поэтому можно зафиксировать любой знак (в данном случае «+») перед $\Phi(t)$ и $\Psi(t)$, лишь бы сохранилось соотношение знаков слагаемых в аргументах косинуса.

Заметим, что если бы огибающие и фазы низкочастотных сигналов определялись не с помощью преобразования Гильберта, то выражение (8.15) было бы неверным, так как только для гильбертовских сигналов поворот фаз всех спектральных со-

ставляющих эквивалентен повороту мгновенной фазы.

Теперь легко видеть, что, сложив полученные сигналы (при передаче на верхней боковой полосе), можно восстановить первичный сигнал $x(t)$ (с точностью до постоянного коэффициента и поворота фаз на $-\psi$) и скомпенсировать зеркальную помеху. При передаче на нижней боковой полосе такой же результат получается при вычитании сигналов (8.15).

Заметим, что и при обычном преобразовании частоты можно воспользоваться компенсационной схемой (рис. 8.7) для подавления зеркальной помехи.

8.6. Фантазии о квадратичном детектировании ОМ сигналов

Пусть в приемном устройстве имеется восстановленная с точной фазой несущая частота, но ее амплитуда не превосходит во много раз уровень детектируемого однополосного сигнала, а детектирование осуществляется путем выделения огибающей суммы боковой полосы и несущей с помощью линейного амплитудного детектора. Такая ситуация может иметь место, например, в «совместимых» однополосных системах, когда в канал посылается боковая полоса и несущая.

Предположим «для простоты» (как это принято во многих учебниках), что модуляция осуществляется одним гармоническим сигналом («чистым тоном») с частотой Ω. Тогда, выбрав отсчет времени так, чтобы начальные фазы равнялись нулю, можно записать несущую частоту как $E_0 \cos \omega_0 t$, а боковую частоту как $E_1 \cos(\omega_0 \pm \Omega) t$. Для нахождения

огибающей суммы боковой и несущей частот

$$s(t) = E_0 [\cos \omega_0 t + (E_1/E_0) \cos (\omega_0 \pm \Omega) t]$$

воспользуемся сопряженным по Гильберту сигналом

$$\hat{s}(t) = E_0 [\sin \omega_0 t + (E_1/E_0) \sin (\omega_0 \pm \Omega) t].$$

Тогда огибающая

$$E(t) = \sqrt{s^2(t) + \hat{s}{}^2(t)} = E_0 [1 + (E_1/E_0)^2 + 2(E_1/E_0) \cos \Omega t]^{1/2}. \quad (8.16)$$

На рис. 8.8 изображена форма огибающей при различных отношениях E_1/E_0. Только при $E_1/E_0 \ll 1$ форма огибающей близка к косинусоидальной. Разлагая (8.16) в ряд Фурье, можно показать, что

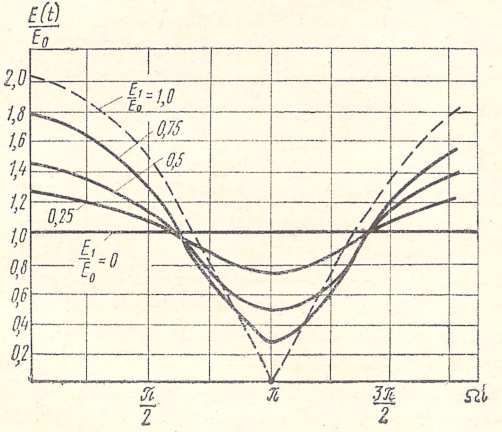


Рис. 8.8. Форма огибающей суммы однополосного гармонического сигнала с несущей

при $E_1/E_0=1/2$ коэффициент 2-й гармоники равен 0,1, а при $E_1/E_0=1$, когда форма огибающей принимает вид, показанный пунктиром на рис. 8.8, коэффициент гармоник достигает 0,25.

В учебнике [12], откуда заимствован рис. 8.8, по этому поводу сказано:

«Совершенно ясно, что в случае линейного детектирования, когда выходное напряжение воспроизводит форму огибающей входного напряжения, отношение E_1/E_0 должно быть значительно меньше единицы, так как в противном случае возникают значительные искажения. При квадратичном же детектировании выходное напряжение, пропорциональное квадрату огибающей,

$$u_{вых}(t) = K[E(t)]^2 = KE_0^2[1 + (E_1/E_0)^2 + 2(E_1/E_0)\cos\Omega$$

воспроизводит частоту модуляции Ω без искажений».

На этом в упомянутом учебнике рассмотрение однополосной модуляции заканчивается и у читателя остается убеждение, что для неискаженного детектирования ОМ сигнала достаточно восстановить точно несущую частоту, сложить несущую с ОМ сигналом и подать сумму на квадратичный детектор. Подобные утверждения содержатся и в ряде других книг.

Все это, конечно, верно, но лишь при условии, что передается один-единственный гармонический сигнал. Если первичный сигнал содержит хотя бы две гармонических составляющих, все коренным образом изменяется. Пусть, например, первичный сигнал равен $E_1 \cos \Omega_1 t + E_2 \cos \Omega_2 t$. Сумма верхней боковой полосы с несущей частотой пропорциональна

$$s(t) = E_0 \cos\omega_0 t + E_1 \cos(\omega_0 + \Omega_1) t + E_2 \cos(\omega_0 + \Omega_2) t.$$

Используя сопряженный сигнал

$$\hat{s}(t) = E_0 \sin \omega_0 t + E_1 \sin (\omega_0 + \Omega_1) t + E_2 \sin (\omega_0 + \Omega_2) t,$$

найдем огибающую

$$E(t) = E_0 [1 + (E_1/E_0)^2 + (E_2/E_0)^2 + 2 (E_1/E_0) \cos \Omega_1 t + \\ + 2 (E_2/E_0) \cos \Omega_2 t + 2 (E_1 E_2/E_0^2) \cos (\Omega_1 - \Omega_2) t]^{1/2}. \quad (8.17)$$

Легко убедиться, что при квадратичном детектировании, хотя и не возникают гармоники частот первичного сигнала, избавиться от комбинационной частоты $\Omega_1 - \Omega_2$ не удается. При более сложном первичном сигнале искажения на выходе квадратичного детектора могут даже превысить искажения на выходе линейного детектора. Для неискаженного детектирования в общем случае необходимо применить синхронный детектор или (если почему-либо желательно выделять огибающую суммы несущей частоты и боковой полосы) обеспечить большое превышение амплитуды несущей над всеми амплитудами боковых частот.

На этом примере можно убедиться, как легко совершаются ошибки, если мы привыкаем «мыслить синусоидой», т. е. представлять первичный сигнал всегда гармонической функцией, о чем уже говорилось ранее (см. стр. 110). Получаемый при этом результат пытаются без проверки отнести и к общему случаю. В приведенной цитате это привело к ошибочному выводу о якобы существующем преимуществе квадратичного детектора при детектировании суммы однополосного сигнала и несущей частоты.

8.7. О требуемой точности восстановления несущей частоты

Известно, что при передаче речи с помощью ОМ не только фаза восстановленной несущей произвольна, но и сама несущая частота может быть восстановлена не абсолютно точно. Это легко понять. Поскольку звуковое восприятие практически не зависит от начальной фазы опорного сигнала в синхронном детекторе, то и медленные изменения этой фазы не должны ощущаться на слух. К этому, в частности, сводится случай, когда значение восстановленной несущей частоты несколько отличается от истинного. Действительно, при однополосной передаче речи и даже музыки расстройка опорного сигнала относительно несущей частоты до 1—2 Гц совершенно не ощущается. При дальнейшем увеличении расстройки постепенно становятся заметнее искажения звука, которые обычно характеризуют как появление хрипа и «металлического» призвука. При расхождении частот в 50—100 Гц эти искажения становятся весьма неприятными, однако разборчивость речи сохраняется достаточно высокой вплоть до расстроек порядка 200 Гц и более. Впрочем передача музыки фактически не достигает цели уже при расстройках в 3—10 Гц.

Рассмотрим, чем определяются приведенные цифры. При восстановлении несущей частоты с погрешностью Δf все спектральные составляющие первичного сигнала оказываются сдвинутыми на Δf. Если бы передавался один тон (гармонический сигнал), то его смещение по частоте на не очень большую величину Δf оказалось бы незамеченным. Известно, что слуховой анализатор человека очень плохо различает абсолютную высоту звука, т. е. его частоту. Лишь отдельные люди, чаще всего

опытные музыканты, умеют это делать с точностью примерно до 5—10%. Это свойство называется абсолютным слухом. Однако практически все люди хорошо различают отношения частот двух одновременно или последовательно звучащих тонов. Так, если частота одного тона точно вдвое больше, чем другого, то их совместное звучание воспринимается как «консонанс» — чистая октава. Если же отношение этих двух частот будет равно не 2, а, скажем, 15/8 или 17/8, то их совместное звучание воспринимается как диссонирующий интервал — большая септима или малая нона. Однако ощущение фальши при восприятии совместного звучания двух звуков наступает уже тогда, когда отношение их частот отклоняется от целочисленного (или выражаемого дробью с малым числителем и знаменателем, например 3/2, 5/4, 6/5) на несколько процентов. На сколько именно? Попробуем оценить этот допуск, не прибегая к трудоемким экспериментам и обследованиям учащихся музыкальных школ.

Вероятно, читателям известно, что музыкальные инструменты с фиксированной настройкой (например, фортепиано, орган, флейта, баян и т. п.) настраиваются так, что только для интервала октавы соблюдается точное целочисленное отношение частот 2 : 1. Для остальных же интервалов эти отношения отклоняются от рациональных, определяемых так называемым натуральным звукорядом. Делается это для того, чтобы получить одинаковые интервалы между соседними звуками звукоряда, который в этом случае называется темперированным. Темперация возможна потому, что средний человеческий слух не отличает темперированных интервалов от натуральных.

Величина интервала между двумя звуками с частотами f_1 и f_2 измеряется величиной $\log (f_2/f_1)$.

Обычно используют основание логарифма 2, так как при этом важнейший интервал — октава — оказывается равным единице. В современном темперированном звукоряде октава разделена на 12 равных (в логарифмическом смысле) частей. 1/12 октавы называется полутоном. Таким образом, полутон является интервалом между двумя соседними звуками темперированного (хроматического) звукоряда.

Заметим, что в натуральном звукоряде все интервалы соответствуют рациональным отношениям частот. Следовательно, величина натурального интервала, т. е. двоичный логарифм этого отношения, не может быть рациональным числом, за исключением того случая, когда отношение частот равно целой степени основания логарифма 2. Но в этом случае интервал равен октаве или нескольким октавам. Таким образом, все натуральные интервалы, отличающиеся от октавы или целого числа октав, измеряются иррациональными числами.

Наоборот, в темперированном строе все интервалы измеряются рациональными числами, кратными 1/12 октавы. Следовательно, отношение частот, образующих темперированный интервал, за исключением октавы, не может быть рациональным. Так, например, интервал квинта в натуральном звукоряде соответствует отношению частот 3/2. Двоичный логарифм этого отношения, т. е. величина этого интервала, если за единицу принята октава, является иррациональным числом 0,58505... В темперированном звукоряде квинта аппроксимируется рациональным числом $7/12=0{,}58333...$ Различие между натуральной и темперированной квинтами составляет, таким образом, около 0,0017 октавы, или немного больше 2 центов (центом называется сотая часть полутона, или 1/1200 октавы). Такая разница не улавливается даже самым изощренным музыкальным слухом.

Наибольшее отличие между темперированным и натуральным интервалами дают увеличенная кварта и уменьшенная квинта, которые в темперированном строе не различаются и составляют половину октавы. Неточность представления соответствующих натуральных интервалов составляет $\pm 0{,}015$ октавы (18 центов). Следовательно, суще-

ствование темперированного строя доказывает, что отклонение интервала на 0,015 практически не влияет на восприятие звука слуховым анализатором.

Пусть теперь в спектре первичного сигнала содержатся частота f_1 и ее вторая гармоника $2f_1$. При детектировании ОМ сигнала с погрешностью Δf они преобразуются соответственно в $f_1+\Delta f$ и $2f_1+\Delta f$, так что вместо интервала 1 (октава) получится

$$i = \log \frac{2f_1+\Delta f}{f_1+\Delta f} = \log \frac{2f_1(1+\Delta f/2f_1)}{f_1(1-\Delta f/f_1)} =$$

$$= 1 + \log[1+\Delta f/2f_1] - \log[1+\Delta f/f_1].$$

Так как $\Delta f \ll f_1$, можно воспользоваться приближением $\log(1+\Delta f/f_1) \approx \frac{\Delta f}{f_1}\log e$, откуда $i \approx 1 + [\Delta f/(2f_1) - \Delta f/f_1] \times \log e \approx 1 - 1{,}45\,\Delta f/(2f_1)$.

Такой сдвиг спектра будет мало заметен на слух, если $1{,}45|\Delta f|/(2f_1) \leqslant 0{,}015$ для всех частот f_1 в спектре первичного сигнала, в том числе и для $f_1 = f_{мин}$ — нижней частоты этого спектра. Тогда допустимая расстройка несущей оказывается равной $|\Delta f|_{доп} \approx 0{,}02\, f_{мин}$. При передаче речи в стандартной полосе частот 300—3400 Гц, $|\Delta f|_{доп} = 0{,}02 \cdot 300 = 6$ Гц. Действительно, такая неточность оказывается практически незаметной. Она была бы столь же незаметной и при передаче музыки, если бы воспроизводилась та же полоса частот. Но обычно при музыкальном вещании стараются передать частоты в более широкой полосе, по крайней мере от 100 Гц. Но в этом случае неточность восстановления несущей будет незаметной только при условии, что она не превышает $0{,}02 \cdot 100 = 2$ Гц.

Более значительные неточности при передаче речи приводят к тому, что вместо гармонических

(целочисленных) первичных соотношений частот будут слышны негармонические. Например, вместо частоты 300 Гц и ее гармоник 600, 900, 1200, 1500 Гц при смещении спектра на 20 Гц получатся частоты 320, 620, 920, 1220, 1520 Гц, ни одна из которых не кратна нижней частоте 320 Гц.

Напомним, что при колебаниях тонкой натянутой струны ее спектр содержит частоты, с большой точностью кратные основной частоте. Если же вместо струны заставить колебаться толстый металлический стержень или рельс, то его спектр будет состоять из частот, существенно отличающихся от кратных. В несколько меньшей степени такая некратность частот обертонов наблюдается при колебании голосовых связок человека, воспалившихся вследствие простуды. Поэтому, как только мы слышим речь со смещенным спектром, в котором слегка ощущается нарушение кратности частот обертонов, мы воспринимаем это как «хрипоту», а при большем смещении спектра возникают ассоциации со звуком, издаваемым металлическим стержнем, и мы говорим о «металлическом тембре» голоса.

Даже при сдвигах спектра на 100—200 Гц разборчивость речи в значительной мере сохраняется, хотя тембр ее становится крайне неестественным и неприятным.

Для объяснения разборчивости следует исходить не из законов слухового восприятия гармонических звуков, а из принципов опознавания членораздельных звуков речи. К сожалению, мы знаем о восприятии речи значительно меньше, чем о восприятии музыки. Однако в первом приближении можно воспользоваться несколько устаревшей, но все же удовлетворительно описывающей факты, формант-

ной теорией. Согласно этой теории звуки речи различаются по размещению максимумов энергетического спектра в определенных полосах частот, называемых формантными областями. Полагая ширину формантных областей порядка 200—300 Гц, легко понять, что сдвиги спектра речи на 100—200 Гц не должны уводить его максимумы за формантные области и поэтому некоторая разборчивость должна сохраняться, что и наблюдается на практике. При сдвигах спектра на 300—400 Гц и более происходит «перепутывание» формант и речь становится неразборчивой.

9

СВЯЗЬ С СОБРАТЬЯМИ ПО РАЗУМУ

9.1. Готовьтесь к межпланетной связи

Внеземные цивилизации ... Что может быть увлекательнее задачи их поиска? И, конечно, этот поиск, а затем и установление связи с ними мыслится сейчас только с помощью электромагнитных волн. В настоящее время в ряде стран, в том числе и в Советском Союзе, проводятся систематические работы по поиску радиосигналов, которые может быть посылаются с неведомых планет далеких галактик. Но обнаружить сигналы — это еще только самое начало. Нужно научиться их понимать и отвечать на них. О трудностях, лежащих на этом пути, хорошо и образно рассказано в [34], и мы не будем на них останавливаться. Предположим, что они уже позади — мы научились хорошо понимать друг друга, договорились о коде, и теперь задача состоит в том, чтобы сообщить разумным существам планеты *X* как можно более подробные сведения о человечестве, его истории, современном состоянии и чаяниях. Сколько на это потребуется времени?

Такой вопрос затрагивается в книге [60], где на с. 266 приводятся следующие рассуждения.

«Известно, что за всю историю человеческой культуры было написано около ста миллионов книг и рукописей. Будем считать (условно), что средний объем одной книги — десять авторских листов. Так как в одном авторском листе содержится, по существующим стандартам, 40 тыс. печатных знаков, то полное количество таких знаков в 100 млн. книг будет $4 \cdot 10^{13}$. Если каждый знак кодировать в двоичной системе ..., полное число знаков двоичного кода, которое должно быть передано, будет порядка $(1 \div 2) \cdot 10^{14}$. Если теперь полоса частот передающегося сигнала будет 1000 МГц, ..., то потребуется 10^5 с, или всего лишь немногим более суток, чтобы передать содержание всего, что когда-либо было написано людьми!».

Такой результат, конечно, поражает читателя. Подумать только, все то, что человечество накапливало за тысячи лет существования цивилизации, можно передать за какие-то сутки. Поневоле хочется преклониться перед достижениями техники связи.

Однако не будем преждевременно радоваться. В приведенной цитате вопрос рассмотрен только с одной стороны — необходимой полосы частот в соответствии с пределом Найквиста (см. гл. 5). Но существуют еще много других проблем, в частности проблема энергетических ресурсов. Попробуем оценить энергию, необходимую для передачи количества информации порядка 10^{14} бит.

9.2. Передача в диапазоне сантиметровых волн

Предположим, что единственной помехой при приеме сигналов является тепловой шум. Пусть приемник находится при температуре жидкого гелия ($\approx$ 4 К) и коэффициент шума его равен единице.

Средняя частота сигнала в соответствии с предыдущим расчетом — порядка $f=10^{10}$ Гц, что соответствует длине волны $\lambda=3$ см. При этом спектральная плотность теплового шума, в Вт/Гц,

$$N_0=\frac{hf\cdot 10^{-7}}{e^{hf/kT}-1}\approx\frac{6{,}5\cdot 10^{-24}}{e^{0{,}118}-1}\approx 5{,}2\cdot 10^{-23}. \qquad (9.1)$$

Здесь $h\approx 6{,}5\cdot 10^{-27}$ эрг·с — постоянная Планка; $k\approx 1{,}38\cdot 10^{-16}$ эрг/град — постоянная Больцмана; T — абсолютная температура.

Поскольку в этих условиях $hf\ll kT$, т. е. $N_0\approx kT$, квантовыми эффектами можно полностью пренебречь, а шум считать белым гауссовским. В этих условиях пропускная способность канала связи на единицу времени, в бит/с, выражается известной формулой Шеннона

$$C=F\log(1+P_{\text{пр}}/N_0F)<(P_{\text{пр}}/N_0)\log e, \qquad (9.2)$$

где F — ширина спектра сигнала; $P_{\text{пр}}$ — мощность сигнала на входе приемника. Неравенство в (9.2) следует из оценки $\ln x\leqslant 1-x$, в которой равенство имеет место только при $x=1$.

Пусть на передачу M бит затрачено t_M секунд. Тогда $M\leqslant Ct_M<(P_{\text{пр}}t_M/N_0)\log e=(E_M/N_0)\log e$, где E_M — энергия сигнала на входе приемника, необходимая для передачи M бит информации, которая в соответствии с полученным неравенством ограничена снизу величиной

$$E_M>MN_0/\log e. \qquad (9.3)$$

Подставляя сюда $M=10^{14}$ бит и $N_0=5{,}2\cdot 10^{-23}$ Вт/Гц, получим, что к приемнику на планете X необходимо подвести энергию сигнала, не меньшую чем $5\cdot 10^{-9}$ Дж. Если при этом $t_M=10^5$ с (приблизительно сутки), то мощность сигнала, подводимая к приемнику, $P_{\text{пр}}>5\cdot 10^{-14}$ Вт.

Перейдем теперь к передатчику. Мощность сигнала на входе приемнике при распространении в свободном пространстве

$$P_{пр} = P_{изл} \lambda^2 G_1 G_2/(4\pi r)^2, \qquad (9.4)$$

где $P_{изл}$ — излучаемая мощность; r — расстояние между передатчиком и приемником; G_1 G_2 — коэффициенты усиления передающей и приемной антенн соответственно.

К тому времени, когда удастся установить контакт с внеземными цивилизациями, радиотехника уйдет далеко вперед. Поэтому примем, что на волне $\lambda = 3$ см коэффициенты усиления антенн $G_1 = G_2 = 10^6$, что превышает сегодняшние возможности. Угол раствора диаграммы направленности такой антенны составляет около 20′.

Для того чтобы вычислить излучаемую мощность, недостает еще одного параметра — расстояния от Земли до планеты X. Здесь можно только гадать, опираясь на высказывания специалистов. В [34] приводится сводка мнений различных ученых, занимавшихся этой проблемой. Среднее расстояние между двумя планетами, на которых существует разумная и достаточно технически развитая жизнь, по их оценкам колеблется от 10 световых лет (по Билсу) до 1000 световых лет (По Хорнеру). По работам других ученых (Саган, Камерон, Опарин, Фесенков, Шкловский) это среднее расстояние ближе к 100 световым годам. Таким образом, существует разброс точек зрения на два порядка. К тому же существует и разброс истинных расстояний относительно среднего.

Предположим, что нам «повезло» и планета X находится всего лишь на расстоянии около 35 световых лет, что составляет $3 \cdot 10^{19}$ см. Подставив эту

величину в ур-ние (9.4), оценим необходимую мощность излучения

$$P_{\text{изл}} = \frac{P_{\text{пр}}(4\pi)^2 r^2}{\lambda^2 G_1 G_2} \approx \frac{5\cdot 10^{-14}\cdot 158\cdot 10^{39}}{9\cdot 10^6\cdot 10^6} \approx 10^{15}\ \text{Вт}.$$

Такую мощность порядка 10^9 МВт могли бы дать 100 тысяч Красноярских ГЭС. Конечно, рассчитывать на получение такой мощности радиоизлучения даже через несколько столетий совершенно не реально. Давайте, однако, помечтаем. Проявим максимум оптимизма и будем полагать, что мощности радиопередатчиков для космической связи к моменту установления контакта с планетой X достигнут 100 МВт[1]. Тогда принимаемый сигнал будет по ф-ле (9.4) иметь мощность около $5\cdot 10^{-21}$. Так как для передачи информации в 10^{14} бит нужно довести до приемника $5\cdot 10^{-9}$ Дж, то на это уйдет время

$$t_M \approx 5\cdot 10^{-9}/5\cdot 10^{-21} = 10^{12}\ \text{с} \approx 30\,000\ \text{лет},$$

т. е. в 3 раза больше тех ста веков, в течение которых человечество добывало и создавало эту информацию!

9.3. Связь в оптическом диапазоне

Нельзя ли, однако, получить более обнадеживающие результаты, выбирая другие параметры системы связи?

Конечно, можно уменьшить требуемую мощность (или требуемое время при данной мощности),

---

[1] Следует заметить, что здесь речь идет о средней, а не о пиковой мощности, подводимой к антенне передатчика. Пиковая мощность в импульсном режиме может намного порядков превосходить среднюю.

снижая температуру приемника. Если снизить ее с 4 К, скажем, до 0,4 К, то все полученные цифры уменьшатся в 10 раз. Но это уменьшение вряд ли можно считать существенным. Надеяться на еще большее снижение температуры не реально.

Глядя на ф-лу (9.4), можно подумать, что к желаемым результатам приведет увеличение длины волны, но это только кажется. При увеличении длины волны уменьшатся коэффициенты усиления антенн (при тех же размерах и сложности) и в результате числитель уменьшится. Но может быть лучше не удлинять, а укорачивать волну?

Действительно, этот путь кажется перспективным. Дело не только в том, что при этом можно увеличить направленность излучения, но и в уменьшении интенсивности теплового шума с повышением частоты, что видно из ф-лы (9.1).

Уменьшим, скажем, длину волны в 50 000 раз, взяв $\lambda = 6 \cdot 10^{-5}$ см, что соответствует видимому свету[1]), и применим в качестве передатчика сверхмощный лазер, который сейчас еще не существует, но несомненно будет построен к моменту организации связи с планетой X. Тогда $f = 5 \cdot 10^{14}$ Гц, и даже при комнатной температуре $T \approx 300$ К $hf/kT \approx \approx 10^3$ и из ф-лы (9.1) получим $N_0 \approx 3 \cdot 10^{-441}$ Вт/Гц, т. е. примерно в 10^{425} раз меньше, чем в предыдущем примере!

Что же, значит во столько же раз уменьшится требуемая мощность сигнала? Тогда для передачи «всей» информации за сутки вместо 10^{15} Вт потребуется всего лишь 10^{-421} Вт? Зачем же тогда говорить о сверхмощном лазере? Ведь самый маломощ-

[1]) Применять еще более короткие волны удастся только в том случае, если передатчик будет вынесен за пределы земной атмосферы, сильно поглощающей ультрафиолетовое излучение.

ный из существующих лазеров обеспечивает несравненно бо́льшую мощность.

К сожалению, это не так. Действительно, при $T=300$ К тепловой шум на частоте $5\cdot10^{14}$ Гц практически полностью отсутствует. Но зато необходимо учитывать ошибки, возникающие из-за квантовой структуры излучения. Дело в том, что для передачи определенного объема информации необходимо, грубо говоря, послать достаточное количество фотонов.

Разберемся в этом примере подробнее. Для идеального квантового канала, в котором не существует тепловых шумов и поглощения фотонов, в [27] получено приближенное выражение пропускной способности (в бит на степень свободы):

$$C \approx \bar{n}\left(1+\ln\frac{1}{\bar{n}}\right)\log e \approx \bar{n}\log\frac{1}{\bar{n}}, \qquad (9.5)$$

где $\bar{n}\ll 1$ — среднее число (по ансамблю сигналов) регистрируемых фотонов[1]).

Для того чтобы найти максимальное количество информации, которое может быть передано в полосе частот F за время t_M, нужно умножить C на Ft_M. Пусть полоса F составляет 20% от несущей f, т. е. $F=10^{14}$ Гц. Для того чтобы передать 10^{14} бит информации, необходимо затратить время t_M и

---

[1]) Выражение (9.5) имеет максимум при $\bar{n}=e^{-1}$, но оно выведено в предположении $\bar{n}\ll 1$. В той области значений $\bar{n}$, для которой оно справедливо, пропускная способность монотонно возрастает при увеличении $\bar{n}$. Заметим, во избежание недоразумений, что, хотя общее число регистрируемых фотонов всегда целое, среднее число фотонов на степень свободы (например, при одной моде за 1 с в полосе 1 Гц) может быть дробным и, в частности, много меньше единицы. Это просто означает, например, что в полосе 1 ГГц фотоны регистрируются значительно реже, чем 10^9 раз в секунду.

число фотонов $\bar{n}t_M$, удовлетворяющие уравнению

$$\bar{n}t_M \log \frac{1}{\bar{n}} F = 10^{14}. \tag{9.6}$$

Выразим $\bar{n}$ через мощность принимаемого сигнала $P_{\text{пр}}$. Если в секунду принимается $\bar{n}F$ фотонов, а энергия каждого из них, как известно, равна hf, то

$$P_{\text{пр}} = \bar{n}Fhf,$$

или

$$\bar{n} = P_{\text{пр}}/(Fhf). \tag{9.7}$$

Излучаемая мощность связана с принимаемой мощностью ф-лой (9.4), которую удобнее для оптического диапазона представить в следующем виде:

$$P_{\text{пр}} = (S_1 S_2 / r^2 \lambda^2)\, \eta\, P_{\text{изл}}, \tag{9.8}$$

где S_1, S_2 — площади передающей и приемной «антенн» соответственно (в данном случае это, по-видимому, площади зеркал телескопов); η — коэффициент, учитывающий поглощение света в оптической системе, атмосфере, приемном устройстве.

Оставаясь на почве, близкой к реальности, будем считать, что диаметры телескопов равны 2 м. Тогда $S_1 = S_2 \approx 3 \cdot 10^4$ см^2. Коэффициент η обычно бывает порядка 10^{-2}.

Будем задаваться различными значениями $\bar{n}$ и вычислять время t_M, необходимое для передачи $M = 10^{14}$ бит информации, и требуемую для этого мощность $P_{\text{изл}}$ по очевидным из предыдущего формулам:

$$t_M = M \Big/ \left(F\bar{n} \log \frac{1}{\bar{n}}\right), \tag{9.9}$$

$$P_{изл} = P_{пр} r^2 \lambda^2 / (S_1 S_2 \eta) = \bar{n} F h f r^2 \lambda^2 / S_1 S_2 \eta. \qquad (9.10)$$

Полученные результаты сведем в табл. 9.1.

ТАБЛИЦА 9.1

$\bar{n}$	$\log \frac{1}{\bar{n}}$	$P_{пр}$, Вт	$P_{изл}$, Вт	t_M, с	$P_{изл}\, t_M$	Примечание t_M
1/128	7	2,35	$2{,}3 \cdot 10^{17}$	7,2	$1{,}7 \cdot 10^{18}$	
10^{-3}	10	0,3	$3 \cdot 10^{16}$	40	$1{,}2 \cdot 10^{18}$	
10^{-6}	20	$3 \cdot 10^{-4}$	$3 \cdot 10^{13}$	20 000	$6 \cdot 10^{17}$	Около шести часов
10^{-9}	30	$3 \cdot 10^{-7}$	$3 \cdot 10^{10}$	$1{,}3 \cdot 10^{7}$	$4{,}5 \cdot 10^{17}$	Примерно 150 суток
10^{-12}	40	$3 \cdot 10^{-16}$	$3 \cdot 10^{7}$	10^{10}	$3 \cdot 10^{17}$	Около 300 лет
10^{-15}	50	$3 \cdot 10^{-13}$	$3 \cdot 10^{4}$	$8 \cdot 10^{12}$	$2{,}5 \cdot 10^{17}$	Около 240 тысяч лет

Примечание. Здесь положено: $f = 5 \cdot 10^{14}$ Гц, $F = 1 \cdot 10^{14}$ Гц, $S_1 = S_2 = 3 \cdot 10^4$ см^2, $r = 3 \cdot 10^{19}$ см и $\eta = 0{,}01$.

Таким образом, для того чтобы передать всю накопленную человечеством информацию даже не за 24 ч, а за 150 суток, потребовалось бы совершенно немыслимая средняя мощность излучения в 30 000 МВт. При средней мощности излучения лазера в 30 кВт передача займет около 240 тысяч лет.

Эти результаты не лучше, чем для волны 3 см. Они свидетельствуют о невозможности передать всю накопленную человечеством информацию за время, существенно меньшее, чем длительность существования цивилизации на Земле.

9.4. Где выход?

Нельзя ли, однако, выбрать параметры системы связи так, чтобы передать 10^{14} бит информации,

если не за сутки, то все же за приемлемый отрезок времени, используя более или менее реальную мощность излучения? Чтобы ответить на этот вопрос, рассмотрим произведение $P_{изл}t_M$, равное общей излученной энергии. Из (9.9) и (9.10) имеем

$$E = P_{изл}t_M = Mhc\lambda r^2/S_1S_2\eta \log(1/\bar{n}), \qquad (9.11)$$

где $c=\lambda f$ — скорость распространения света.

Эта энергия зависит от $\bar{n}$, которая, в свою очередь, определяется мощностью, длиной волны и шириной спектра [1]. Но так как $\bar{n}$ входит в аргумент логарифма, то его изменения мало влияют на требуемую энергию. Это подтверждается предыдущей таблицей, так как при уменьшении $\bar{n}$ в 10^{12} раз произведение $P_{изл}t_M$ снижается лишь в 5 раз. Если отвлечься от этой слабой зависимости, то из (9.11) следует, что уменьшения необходимой энергии можно добиться только, увеличивая S_1 и S_2 и η и уменьшая λ.

Что касается площадей антенн S, то трудно надеяться на возможность их существенного увеличения (под существенным увеличением здесь понимается увеличение на несколько десятичных порядков). Коэффициент η по своему смыслу меньше единицы. Так что остается единственный реальный путь — укорочение длины волны λ.

Если задаться целью передать 10^{14} бит хотя бы за 30 лет (т. е. $\approx 10^9$ с), излучая при этом среднюю мощность, не превышающую 10^5 Вт, то произведение $P_{изл}t_M$ составит 10^{14} Дж. В рассмотренном выше примере при $\lambda=6\cdot10^{-5}$ см это произведение

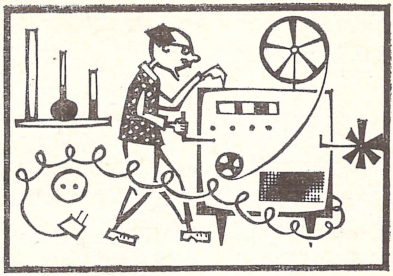
[1] Этим квантовые каналы, между прочим, отличаются от идеальных «классических», в которых минимальная энергия, которая должна быть доведена до приемника, определяется лишь объемом информации и уровнем аддитивного шума и не зависит от мощности сигнала.

было порядка 10^{18} Дж. Следовательно, можно, в принципе, достичь цели, укоротив длину волны примерно в 10^4 раз, т. е. выбрав $\lambda \approx 10^{-8}$ см, или 1 А. Это значение лежит в диапазоне жестких рентгеновских волн.

Итак, если мы когда-нибудь научимся излучать (и модулировать) волны длиной в один ангстрем с мощностью порядка 100 кВт да еще сумеем фокусировать их «антеннами» с диаметром в 3 м, а при передаче поглощаться будет не более 99%, то мы сможем за 30 лет передать на планету X всю накопленную человечеством информацию. Трудно все же поверить, что мы научимся все это делать за несколько ближайших столетий.

Как же быть?

Очень просто. Не будем стараться передать все, что было когда-либо написано. Ограничимся, скажем, передачей Большой советской энциклопедии, содержащей «всего лишь» около 10^9 бит. Если мы будем обладать «сверхлазером» с мощностью 10 кВт на волне, $\lambda = 6 \cdot 10^{-5}$ см и телескопом с $S = 3 \cdot 10^4$ см^2, то мы сможем это сделать за 6 лет или за 60 лет при более реальном значении мощности в 1 кВт. Но прежде всего нужно найти эту планету X с разумными жителями и научиться понимать друг друга.



10

ЛУЧШЕ НАИЛУЧШЕГО

10.1. Изобретатели «вечного двигателя»

За последние 20—25 лет автору приходится время от времени консультировать различных изобретателей систем связи. И примерно каждые 3—4 года среди них появляется очередной изобретатель «вечного двигателя» в связи. Под этим названием я подразумеваю различные способы передачи и приема сигналов с помехоустойчивостью, превышающей потенциальную, т. е. ту, которую можно теоретически получить при оптимальном приеме в данных условиях.

Среди этих изобретателей встречались разные люди. Они по-разному реагировали на критику. Некоторые (очень немногие) сразу понимали свою ошибку и отказывались от своих заблуждений. Другие, разобравшись в слабых местах своих рассуждений, все же не хотели отбросить основу своей идеи и упорно пытались «подправить» и «обосновать» ее. Третьи просто не хотели ничего слушать и объясняли все возражения тем, что их оппоненты — рутинеры, отставшие от жизни, не понимающие новых идей и цепляющиеся за «устаревшие»

представления о потенциальной помехоустойчивости. Нередко в подтверждение своих идей они приводят неизвестно как и где полученные «экспериментальные результаты». Впрочем некоторые из авторов «вечного двигателя» — очень приятные люди и в общем весьма квалифицированные специалисты, отлично владеющие методами теории связи, но упорно допускающие ту или иную ошибку при анализе своей системы.

В ряде случаев значительно легче разобраться в сущности сделанных ошибок, чем в психологии «изобретателей».

В качестве примера можно привести мою переписку с одним инженером, длившуюся более двух лет, в которой речь шла о двоичной системе связи, якобы позволяющей при некогерентном приеме ортогональных сигналов получить такую вероятность ошибки, которая теоретически возможна лишь при когерентном приеме противоположных сигналов. Сущность этой системы и ее многочисленных вариантов (в каждом новом письме вносились новые «усовершенствования») излагать не будем. В ней нет ничего поучительного, и ее опровержение сводилось к отысканию тривиальных ошибок в выкладках. Так, в третьем или четвертом письме предлагалась модификация решающей схемы, в которой отношение сигнал/помеха оказывалось пропорциональным $1/(N^2_1 - N^2_2)$, где N_1 и N_2 — интенсивности шумов в двух ветвях решающей схемы. Повторив сделанные в письме выкладки, я сразу обнаружил, что в этой формуле ошибочно записан знак, так что вместо $N^2_1 - N^2_2$ в знаменателе должно быть $N^2_1 + N^2_2$ и никакой сверхъестественной помехоустойчивости не получается. Указав в своем ответе на эту и еще на одну ошибку, я закончил его следующей фразой:

«Я очень боюсь, что, прочтя это письмо и убедившись в его справедливости, Вы начнете «подправлять» и «совершенствовать» свой алгоритм, стараясь все-таки превзойти потенциальную помехоустойчивость. Не уподобляйтесь тем изобретателям вечного двигателя, которые считали, что у них недоработана только одна небольшая деталь, а когда она будет улучшена, вечный двигатель заработает».

В ответ я получил письмо, автор которого категорически отрицал свою явную ошибку. Он писал:

«Расхождения в знаках в формулах (8) и (9) объясняются тем, что при выкладках Вы не имели возможности учесть временную структуру процессов в ветвях графа: разъяснение деталей технологии не входило в задачи материала».

Другими словами, главная «шестеренка» вечного двигателя оказалась засекреченной — тут уж ничего возразить нельзя.

Справедливости ради нужно сказать, что примерно через три года мой корреспондент докладывал о своем алгоритме на одной научной конференции и указанные ошибки были им исправлены. Но, конечно, никакого превышения потенциальной помехоустойчивости уже не было. При обсуждении доклада выяснилось, что этот алгоритм ничем не лучше известных.

В чем же корень всех, заранее обреченных на неуспех, попыток преодолеть законы природы?

В первую очередь он заключается в отсутствии самокритичного отношения к своим результатам. К сожалению, далеко не все преподаватели вузов воспитывают в своих учениках привычку быть всегда самым строгим критиком своей работы и соблюдать незыблемое правило — после получения какого-либо результата прежде всего стараться его опровергнуть, и лишь после того, как он выдержал все испытания, выставлять его на всеобщее обозрение.

Другим корнем является весьма похвальное желание сделать великое открытие, принести пользу человечеству, ну и за одно прославиться. Если это желание не сдерживается самокритикой, то часто оно приводит к плачевным результатам. Хуже всего, когда автор уже свыкся с мыслью, что он изобрел нечто небывалое, и ему очень трудно с ней расстаться. Так рождаются «непризнанные гении», которые принимают всякую критику как проявление зависти.

И, наконец, в отдельных случаях немалую роль играют недостаточная техническая грамотность и

непонимание основ статистической теории связи. Иногда «изобретатели» даже не догадываются о том, что их результаты противоречат теории. Чаще они пытаются объяснить эти противоречия тем, что они глубже и правильнее понимают сущность помехоустойчивости, чем их оппоненты, и могут внести «поправки» в существующую теорию. Приходилось слышать следующие пояснения:

«Котельников в теории потенциальной помехоустойчивости рассматривал только линейные методы обработки сигнала, а у меня — нелинейная схема, и поэтому я могу обеспечить более высокую помехоустойчивость»;

«Котельников исходил только из представления сигнала как функции времени или из разложения этой функции в ряд Фурье, если же учесть различные другие свойства сигналов, то можно найти резервы для повышения помехоустойчивости»;

«Котельников рассматривал только белый шум, но в природе белого шума не бывает, и поэтому вероятность ошибки может быть сколь угодно малой, если придумать хорошую систему»;

«Теория потенциальной помехоустойчивости давно опровергнута Шенноном, который доказал, что информацию можно передавать со сколь угодно малой вероятностью ошибки, если скорость передачи меньше пропускной способности канала» и т. п.

Все эти доводы, конечно, несостоятельны. При построении теории потенциальной помехоустойчивости В. А. Котельников и продолжатели его работы не постулировали линейности методов обработки сигнала. Линейная система оказывается оптимальной только для гауссовской помехи при полностью известном сигнале. В других случаях оптимальная схема получается нелинейной. Точно

так же никаких ограничений в представлении сигнала в теории Котельникова нет. Запись принимаемого сигнала в виде функции времени содержит всю доступную информацию о нем, и никакие преобразования этой записи не смогут выявить новой информации, а следовательно, и не позволят получить результаты, лучшие, чем в синтезированной по выбранному критерию оптимальной схеме [1].

Тот факт, что в работе В. А. Котельникова [23] рассматривается в основном только помеха в виде белого гауссовского шума, вовсе не говорит о возможности получения сколь угодно малой вероятности ошибок при небелом шуме. Нижняя граница для вероятности ошибок при небелом гауссовском шуме определялась во многих работах других авторов, например в [44]. Эта граница становится равной нулю в так называемых сингулярных случаях, например, когда в некоторой частотной области, занятой сигналом, помеха отсутствует. Однако в реальных ситуациях сингулярные случаи возникнуть не могут, хотя бы вследствие всегда присутствующего теплового шума, а также квантовых закономерностей. Сингулярности возникают только «на бумаге», когда принятая математическая модель канала связи физически нереализуема и, следовательно, неправильно описывает реальный канал. К тому же изобретатели «сверхоптимально-

---

[1] Конечно, используя дополнительную информацию, можно снизить вероятность ошибки. Например, если вместо скалярной функции — напряжения на входе приемника — задать пространственную картину электромагнитного поля вблизи приемной антенны, то возникают новые возможности — использования поляризации поля, осуществления пространственно-разнесенного приема и т. д. Ничего похожего не будет, если, например, помимо сигнала, задать его производную, поскольку она новой информации не содержит.

го приема», ссылающиеся на окрашиваемость шума, в своих работах рассматривают обычно несингулярную модель, но получают вероятность ошибки если не сколько угодно малую, то все же меньшую, чем это возможно в оптимальной решающей схеме.

Совершенно незаконным является противопоставление теорий Котельникова и Шоннона. В теории потенциальной помехоустойчивости Котельникова считается заданным конечный ансамбль сигналов и отыскивается вероятность ошибок при оптимальном приеме этих сигналов по критерию максимальной апостериорной вероятности, тогда как в теории информации Шеннона задан только канал, а сигналы выбираются оптимальным образом в процессе кодирования. Уменьшение вероятности ошибок имеет место при росте длины кодируемого сообщения, когда растет также и число реализаций сигнала. Впрочем и в [23] показано, что с увеличением основания кода вероятность ошибки (пересчитанная на двоичный символ) уменьшается.

После этих общих соображений перейдем к знакомству с некоторыми конкретными способами ниспровержения теории связи. Ограничимся только тремя типичными примерами.

10.2. Широкополосная частотная манипуляция

Эта «идея» в законченном виде была обнаружена в рукописи статьи, которая, к счастью, не была опубликована. Автором ее была аспирантка Пивоварова, и, как потом выяснилось, в статье излагалось одно из основных положений подготовленной ею диссертации (все фамилии «изобретателей» в этой главе изменены).

Суть дела состоит в следующем. Известно, что при некогерентном приеме двоичных сигналов с частотной манипуляцией (ЧМ) на фоне гауссовского шума с равномерным спектром обычно применяют пару фильтров, разделяющих две реализации сигнала, и дифференциальный детектор (рис. 10.1). Решение принимают по знаку напря-

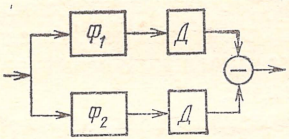


Рис. 10.1. Прием частотноманипулированных сигналов разделительными фильтрами:
Ф — фильтр; *Д* — детектор

Рис. 10.2. Прием ЧМ сигналов по мгновенной частоте:
ЧД — частотный детектор; *ФНЧ* — фильтр нижних частот

жения в моменты отсчета на выходе детектора. При достаточно большой разности частот сигналов и достаточно широких, но неперекрывающихся полосах пропускания фильтров вероятность ошибки

$$p = 0{,}5e^{-q^2/2}, \qquad (10.1)$$

где q^2 — отношение мощности сигнала к мощности помехи на выходе фильтра.

Если фильтры согласованы с сигналами, а последние взаимно ортогональны, то в (10.1) величина q^2 принимает максимально возможное значение, равное отношению энергии сигнала на входе приемника к спектральной плотности шума, которое принято обозначать h^2. В этом случае схема оказывается оптимальной по критерию максимума правдоподобия. Таким образом, $0{,}5\,e^{-h^2/2}$ представ-

ляет наименьшую вероятность ошибки при некогерентном приеме двоичных ЧМ сигналов. К ней можно приближаться, используя и некоторые другие схемы, например прием по мгновенной частоте с последетекторным фильтром (рис. 10.2) (см., например, [46], с. 262—291).

Тов. Пивоварова обнаружила, что в этих рассуждениях имеется неточность. При детектировании ЧМ сигнала, как известно, отношение мощности сигнала к мощности помехи возрастает в $3m^2/\Pi^2$ раз, где m — индекс частотной модуляции, Π — пикфактор модулирующего сигнала (см., например, [15]). В случае частотной манипуляции гармонического переносчика пикфактор равен 1, а за индекс модуляции можно в первом приближении принять $k=\Delta f/F_0$, где Δf — девиация частоты, равная половине разности частот двух реализаций («нажатия» и «отжатия»), F_0 — частота манипуляции, т. е. величина, обратная длительности двух элементов сигнала, $1/(2T)$.

Правда, этот выигрыш имеет место только при условии, что отношение сигнал/помеха на входе приемника превышает некоторое пороговое значение, которое при ЧМ близко к 3. Но это на практике всегда выполняется, если качество связи хотя бы приближается к удовлетворительному.

Исходя из этого аспирантка Пивоварова предложила увеличивать помехоустойчивость, применяя ЧМ с большой девиацией. Пусть, например, передача идет со скоростью 100 Бод, т. е. $F_0=$ $=50$ Гц, а отношение энергии элемента сигнала к спектральной плотности помехи на входе приемника $h^2=10$. Применяя схему рис. 10.1 с согласованными фильтрами и выбирая $\Delta f=kF_0$, где k — любое целое число (это обеспечивает ортогональ-

ность сигналов), получим вероятность ошибок по ф-ле (10.1)

$$p = 0{,}5e^{-5} \approx 3{,}3 \cdot 10^{-3}$$

независимо от значения k.

Тов. Пивоварова предлагает использовать достаточно большое значение $k>1$, например, хотя бы $k=2$ [1], и вести прием по схеме рис. 10.2. После частотного детектора и фильтра нижних частот получим прямоугольные видеоимпульсы с отношением сигнал/помеха в $3m^2/\Pi^2=3k^2=12$ раз больше, чем h^2, т. е. равное $q^2_{\text{вых}}=120$. Полагая в первом приближении помеху на выходе детектора также гауссовской (что не может существенно повлиять на результат), будем вычислять вероятность ошибки при приеме двухполярных видеоимпульсов на фоне гауссовской помехи при отношении мощностей сигнал/помеха $q^2=120$. Она равна

$$p = 0{,}5\left[1 - \Phi\left(\sqrt{2}q\right)\right], \qquad (10.2)$$

где $\Phi(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\xi^2/2}\, d\xi$ — функция Крампа.

Воспользовавшись известным неравенством $\Phi(x) \geqslant 1 - e^{-x^2/2} d\xi$, справедливым при $x \geqslant 0$, получим $p \leqslant 0{,}5\, e^{-q^2} = 0{,}5\, e^{-120} < 10^{-51}$, т. е. практически прием будет безошибочным!

Из этого делался вывод, что для приема дискретных ЧМ сигналов нужно пользоваться только схемой рис. 10.2 (прием по мгновенной частоте), так как, выбирая при этом достаточно большие индексы модуляции, можно получить практически

[1] Большее значение k в данном примере брать не следует, так как иначе отношение мощности сигнала к полной мощности помехи на входе приемника, которое, очевидно, равно $q^2=h^2/k$, окажется ниже порогового.

безошибочный прием в условиях, когда более распространенная схема рис. 10.1 дает вероятность ошибок порядка 10^{-3}, превышающую допустимое значение в системах передачи данных. В статье делался также вывод, что система с ЧМ при приеме по схеме рис. 10.2 значительно более помехоустойчива, чем с фазовой (ФМ) или фазоразностной (ФРМ) манипуляцией. Правда, в статье, насколько помнится, не подчеркивалось, что полученная помехоустойчивость выше потенциальной.

На этом месте рекомендуем читателю отложить книгу и попытаться сообразить, где кроется ошибка в рассуждениях аспирантки Пивоваровой. К этом вопросу мы еще вернемся, а пока займемся другим предложением, позволяющим, по словам его автора, существенно повысить помехоустойчивость радиотелеграфной связи.

10.3. Интервальная манипуляция

Доцент одного вуза (назовем его Дьяковский), долгое время занимавшийся радиолокацией, обратил внимание на то, что достаточно надежное определение дальности до цели по времени прихода радиолокационного импульса обеспечивается при столь низких отношениях мощности сигнал/помеха, при которых радиотелеграфный прием практически невозможен. Размышляя об этом, он пришел к выводу, что применяемые в настоящее время радиотелеграфные сигналы выбраны неудачно. Основной их порок он видел в использовании двоичного кода, придуманного, как он говорил, «механиком Бодо, ничего не понимавшим в вопросах передачи информации». Взамен двоичного кода Дьяковский предложил неравномерный интервальный код, при котором информация о передаваемой

букве заложена в длительности интервала между двумя соседними импульсами аналогично тому, как в радиолокации дальность до цели определяется длиной интервала между переданным и принятым отраженным импульсом. Каждая буква (кроме первой) передается одним импульсом; интервал отсчитывается от импульса предыдущей буквы (рис. 10.3, где для примера показана последовательность импульсов при передаче слова «едва»).

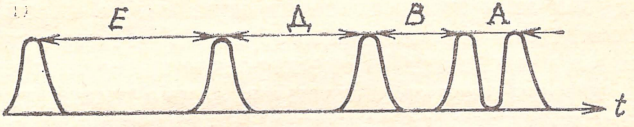


Рис. 10.3. Пояснение неравномерного интервального кода

С этим предложением тов. Дьяковский выступал в течение многих лет на различных научных конференциях и семинарах. Постепенно от сырой идеи он подошел к построению некоторой теории, в которой сравнивались различные параметры существующих и предлагаемого способов радиотелеграфии. Эту теорию можно изложить следующим образом.

В существующих телеграфных системах каждая буква передается с помощью пяти или, в стартстопном варианте, семи тактовых интервалов. Длительность тактового интервала обозначим T. Ширину спектра сигнала F можно оценить в первом приближении величиной k/T, где k — коэффициент, зависящий от формы сигнала и чаще всего принимающий значение между 1,5 и 5.

Пусть в системе, предложенной Дьяковским, используются импульсы сигнала такой же дли-

тельности и формы и, следовательно, с такой же шириной спектра. Интервалы между буквами кратны $T/2$. Поскольку время, затрачиваемое на передачу различных букв, неодинаково, оценим среднюю длительность буквы. Если бы 32 буквы передавались с равной вероятностью, то средняя длительность равнялась бы

$$T_{\text{б.ср}} = T + \frac{T}{2 \cdot 32}(1 + 2 + \cdots + 32) = 8{,}75T. \quad (10.3)$$

Однако если при построении кода учесть неравновероятность использования букв в осмысленном тексте и назначить самые короткие интервалы часто встречающимся буквам, а самые длинные — редким буквам, то величина $T_{\text{б.ср}}$ уменьшится примерно до $6T$, т. е. будет даже меньше длительности буквы в современной стартстопной системе.

Будем считать, что в канале присутствует аддитивный белый гауссовский шум со спектральной плотностью N_0. Сигнал вместе с шумом проходит через согласованный с сигналом фильтр (фильтр Норса) [1]. Решение о переданной букве принимается путем измерения момента пересечения огибающей сигнала некоторого порогового уровня. В отсутствие помех этот момент точно соответствовал бы моменту передачи импульсов, так что все ин-

---

[1] В некоторых своих выступлениях тов. Дьяковский утверждал, что одним из преимуществ его системы является возможность осуществить оптимальную (согласованную) фильтрацию, которая якобы невозможна в системах с двоичным кодом, поскольку длительность отклика фильтра на сигнал равна $2T$, а два импульса сигнала могут присутствовать рядом. Читателю, надеемся, ясно, что это утверждение не имеет под собой никакой почвы. Согласованный фильтр вовсе не противопоказан и при двоичном коде, поскольку к моменту отсчета отклика фильтра на второй импульс его реакция на предыдущий полностью затухает.

тервалы между импульсами измерялись бы без искажений и решение о переданной букве принималось бы безошибочно. Помеха, однако, искажает ход огибающей, и момент ее пересечения с пороговым уровнем может наступить несколько раньше или несколько позже, чем в отсутствие помех (рис. 10.4).

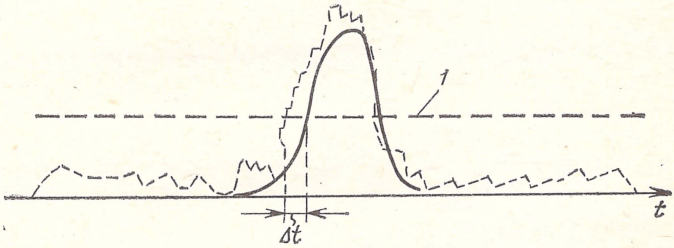


Рис. 10.4. Огибающая принимаемого импульса (*1* — пороговый уровень):
—— без помех; — — — с помехой

В [23], где дан полный анализ этой задачи, показано, что отклонение Δt измеренного момента прихода импульса от истинного является с хорошим приближением гауссовской случайной величиной с нулевым математическим ожиданием. Его дсиперсия σ^2 определяется для идеального приемника [1] ф-лой (7.13) упомянутой работы, которая в

[1] Идеальный приемник анализирует весь ход огибающей. Для реального приемника, отмечающего момент пересечения переднего фронта огибающей с пороговым уровнем, дисперсия увеличивается примерно в 1,8 раза, если пороговый уровень выбран оптимальным. Однако если в приемнике отмечаются пересечения обоих фронтов с пороговым уровнем и берется их средняя величина, то дисперсия оценки практически оказывается такой же, как и в идеальном приемнике.

наших обозначениях имеет вид [1] $\sigma^2 = 6N_0/(2\pi F)^2 Q^2 = 6/(2\pi hF)^2$, где Q^2 — энергия импульса; $h^2 = Q^2/N_0$.

Ошибочное решение о принятой букве произойдет в том случае, когда $|\Delta t|$ превзойдет половину шага в шкале интервального кода, которую мы приняли равной $T/4$, т. е. вероятность ошибочного решения

$$P_{\text{ош}} = P\{\Delta t > T/4\} = \int_{-\infty}^{-T/4} w(\Delta t)\, d\Delta t + \int_{T/4}^{\infty} w(\Delta t)\, d\Delta t =$$

$$= \frac{2}{\sqrt{2\pi}\sigma} \int_{T/4}^{\infty} e^{-x^2/(2\sigma^2)}\, dx = 1 - \Phi\left(\frac{T}{4\sigma}\right) =$$

$$= 1 - \Phi\left(\frac{\pi hFT}{2\sqrt{6}}\right) \approx 1 - \Phi(0{,}64kh), \qquad (10.4)$$

где $k = FT$.

Сравним эту вероятность ошибки с той, которая может быть получена при двоичном коде, например, при наиболее помехоустойчивой фазовой манипуляции с когерентным приемом. При той же пиковой мощности сигнала вероятность ошибочного приема двоичного символа

$$p = 0{,}5[1 - \Phi(\sqrt{2}h)], \qquad (10.5)$$

Легко видеть, что при достаточно большом k вероятность ошибки, определяемая по ф-ле (10.4), меньше, чем по ф-ле (10.5). Например, при $h=2$ и $k=4$ ф-ла (10.4) дает $P_{\text{ош}} \approx 10^{-7}$, тогда как по ф-ле (10.5) — $p \approx 2{,}5 \cdot 10^{-3}$.

---

[1] В работе Котельникова [23] фигурирует дисперсия δ^2 оценки безразмерного параметра λ, равного в данном случае $\Delta t/0{,}25T$, а через σ^2 обозначена величина N_0.

Таким образом, уже на этом этапе рассуждений виден заметный выигрыш, даваемый предложением Дьяковского по сравнению с двоичной фазовой манипуляцией.

Однако сравнивать величины $P_{ош}$ из (12.4) и p из (12.5) не вполне корректно. Первая из этих величин дает вероятность ошибочного приема буквы (из алфавита в 32 буквы), а вторая — двоичного символа. Буква при передаче двоичным безызбыточным кодом будет принята ошибочно, если хотя бы один из пяти входящих в ее кодовую комбинацию символов окажется неверным. Вероятность этого для рассматриваемой модели канала, в которой ошибки происходят независимо друг от друга,

$$P_{ош} = 1 - (1-p)^5 = 1 - \frac{1}{32}\left[1 + \Phi\left(\sqrt{2h}\right)\right]^5, \quad (10.6)$$

что дает при $h=2$ для системы двоичной ФМ $P_{ош} \approx 0{,}0125$.

Итак, при одинаковой пиковой мощности сигнала предложенная система позволяет уменьшить вероятность ошибок примерно в 100 тысяч раз по сравнению с двоичной системой ФМ. Для того чтобы получить вероятность ошибочного приема буквы в двоичной системе ФМ порядка 10^{-7}, потребовалось бы значение h около 4, т. е. пиковую мощность сигнала нужно было бы увеличить в 4 раза. Для иллюстрации на рис. 10.5 показаны зависимости $P_{ош}$ от h, вычисленные по ф-лам (10.6) и (10.4) при $k=4$.

Но и это еще не все. Довольно часто приходится сравнивать системы связи не по пиковой, а по средней мощности. При этом выигрыш системы Дьяковского возрастает еще примерно в 6 раз, так как в системе ФМ пауз нет, а в предложенной си-

стеме в среднем один импульс длительности T приходится на интервал времени $6T$.

Наконец, из (10.4) видно, что, увеличивая коэффициент k, т. е. расширяя спектр сигнала и приближая форму его огибающей к прямоугольной,

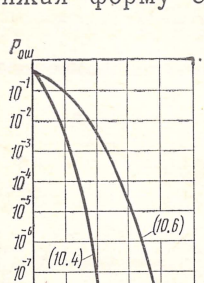


Рис. 10.5. Вероятности ошибочного приема буквы для системы Дьяковского [по ф-ле (10.4)] и для ФМ

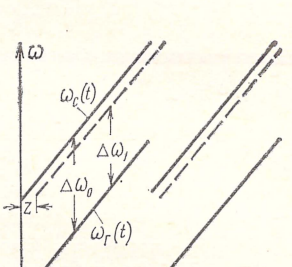


Рис. 10.6. Изменения мгновенной частоты ЛЧМ сигнала и гетеродина:
— — — частота сигнала после линии задержки

можно еще уменьшать вероятность ошибочного приема, т. е. увеличивать энергетический выигрыш системы Дьяковского [1].

После всего сказанного приходится удивляться, почему до сих пор система Дьяковского не реализована, автор ее не стал Нобелевским лауреатом, а все его выступления перед квалифицированной

[1] В ф-ле (10.5) коэффициент k не фигурирует. Известно, что вероятность ошибки в двоичной системе с противоположными сигналами при оптимальном приеме в канале с гауссовским аддитивным белым шумом не зависит от формы или спектра сигнала и определяется только отношением энергии сигнала к спектральной плотности шума.

аудиторией встречались весьма иронически. Ответ на этот вопрос мы пока отложим и перейдем к рассмотрению еще одного предложения, позволяющего превысить потенциальную помехоустойчивость при передаче дискретных сообщений.

10.4. «Сверхоптимальный» прием

В отличие от многих других изобретателей «вечного двигателя» в связи, тов. Маляров весьма эрудирован, хорошо знает теорию потенциальной помехоустойчивости и владеет математическим аппаратом. К тому же он безусловно талантливый инженер и очень приятный собеседник. Его предложение состоит в использовании сигналов с линейной частотной модуляцией (ЛЧМ) при специальном «многоканальном» способе обработки принятого сигнала.

Сигнал ЛЧМ можно записать в следующем виде:

$$s(t) = U_0 \cos(\omega_0 t + \pi\gamma t^2 + \varphi^2), \quad 0 \leqslant t \leqslant T, \qquad (10.7)$$

где $\gamma = \Delta f/T$, Δf — ширина полосы, в которой изменяется мгновенная частота

$$\omega(t) = \frac{d}{dt}(\omega_0 t + \pi\gamma t^2 + \varphi) = \omega_0 + 2\pi\gamma t. \qquad (10.8)$$

Из (10.8) видно, что мгновенная частота изменяется по линейному закону (рис. 10.6). Выберем значение γ настолько большим, что $\Delta f = \gamma T \gg 1/T$.

Двоичную манипуляцию такого сигнала можно осуществить, например, изменяя исходную частоту ω_0, или направление изменения частоты (знак γ), или начальную фазу φ и т. п. Для упрощения примем, что используется амплитудная манипуляция (АМ), т. е. сигнал $s(t)$ либо присутствует, ли-

бо отсутствует, так что задача сводится к обнаружению сигнала $s(t)$ на фоне гауссовского шума.

Один из возможных методов приема заключается в использовании ЛЧМ гетеродина, генерирующего напряжение

$$u_г(t) = U_г \cos(\omega_г t + \pi\gamma t^2 + \psi), \qquad (10.9)$$

мгновенная частота которого изменяется по тому же закону, но со сдвигом на величину $\Delta\omega_0 = \omega_0 - \omega_г$, как показано на рис. 10.6, причем частобы выбраны так, что $\Delta\omega_0 \ll \omega_г < \omega_0$.

Перемножив принятый сигнал на напряжение гетеродина (рис. 10.7), получим в точке a

$$s(t)\,u_г(t) = \frac{1}{2} U_0 U_г \{\cos[\Delta\omega_0 t + \varphi - \psi] + \cos[(\omega_0 + \omega_г)t + 2\pi\gamma t^2 + \varphi + \psi]\}.$$

Спектр первого члена в фигурных скобках состоит из одной составляющей на частоте $\Delta\omega_0$, а основная часть спектра второго члена распределена в полосе частот $(\omega_0 + \omega_г) \div (\omega_0 + \omega_г + 4\pi\Delta f)$ (см. гл. 7). Таким образом, в точке a сигнал представляет собой отрезок гармонического колебания на частоте $\Delta\omega_0$, другими словами, обычный сигнал с АМ плюс далеко отстоящий по спектру второй член.

Фильтр согласован с полученным гармоническим сигналом на частоте $\Delta\omega_0$, и его эквивалентная полоса пропускания равна $1/T$. Если на входе приемника присутствует аддитивный гауссовский белый шум со спектральной плотностью N_0, то и в точке a шум будет гауссовским, так как операция умножения на напряжение гетеродина является линейной. Отношение мощности сигнала к мощности шума в точке b равно отношению энергии

сигнала E_s к спектральной плотности шума N_0 на входе приемника [1], которое обозначим h^2. Таким образом, по величине напряжения на выходе фильтра в момент отсчета можно судить о том, присутствует или отсутствует сигнал. Вероятность ошибки при этом определяется отношением сигнал/шум h^2.

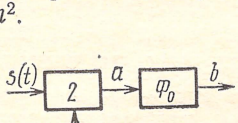


Рис. 10.7. Схема приема ЛЧМ сигнала:
1 — ЛЧМ гетеродин; *2* — перемножитель; *Ф* — согласованный фильтр

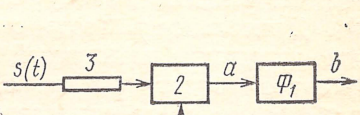


Рис. 10.8. Прием задержанного ЛЧМ сигнала:
1 — ЛЧМ гетеродин; *2* — перемножитель; *3* — линия задержки; *Ф* — согласованный фильтр

Рассмотрим теперь, что произойдет, если сигнал перед подачей на перемножитель задержать на время $z \approx 1/\Delta f$ (откуда следует $z \ll T$). Тогда в точке a рис. 10.8 получим напряжение

$$s(t-z)\,u_г(z) = U_0 U_г \{\cos[\omega_0 (t-z) + \pi\gamma (t-z)^2 + \varphi] \cos[\omega_г t + \pi\gamma t^2 + \psi]\} = \frac{1}{2} U_0 U_г \cos[(\omega_0 - \omega_г) t - 2\pi\gamma z t - \omega_0 z + \pi\gamma z^2 + \varphi - \psi] + X(t), \qquad (10.10)$$

[1] В этом месте в рассуждениях Малярова имеется неточность, так как после перемножения в полосе пропускания фильтра Φ будет присутствовать не только та часть шума, которая возникла от полосы входного шума, лежащей вокруг частоты ω_0, но и «зеркальная» полоса, лежащая вокруг частоты $2\omega_г - \omega_0$, так что в действительности отношение сигнал/шум в точке b будет равно $h^2/2$. Однако это обстоятельство несущественно для последующего, к тому же, в принципе, зеркальную помеху можно подавить фильтром на входе приемника.

где через $X(t)$ обозначена составляющая «суммарной частоты» со спектром, расположенным вблизи частоты $\omega_0+\omega_г$, которая не пройдет через фильтр Φ_1 и поэтому не представляет интереса. Полезный член представляет собой отрезок гармонического колебания на частоте $\Delta\omega_1=\omega_0-\omega_г-2\pi\gamma z=\Delta\omega_0-2\pi\gamma z$. Именно на эту частоту настроен фильтр Φ_1. Длительность этого отрезка равна $T-z$, так что отношение мощностей сигнал/шум на выходе фильтра будет равно $h^2_1=E_c(1-z/T)/N_0$. При $z\ll T$ h_1 почти не отличается от h, так что вероятность ошибки в схеме рис. 10.8 практически такая же, как и в схеме рис. 10.7 [1].

Объединим теперь обе схемы (рис. 10.9). Здесь сигнал поступает на два перемножителя, причем на один из них через линию задержки. На выходах фильтров Φ_0 и Φ_1 получим два сигнала, одинаковых, если не считать сдвига по частоте на $\gamma z\approx 1/T$. Шумы на выходе этих фильтров лежат в непересекающихся полосах частот. Поэтому они некоррелированны. По существу, здесь получились две ветви разнесенного приема с коррелированными сигналами и некоррелированными шумами. Осуществив некогерентное сложение этих сигналов, получим на выходе сумматора отношение сигнал/шум, практически равное $2E_c/N_0=2h^2$. Соответственно вероятность ошибки при обнаружении сигнала уменьшится.

Дальнейшее обобщение очевидно. Используем линию задержки с k отводами через интервалы времени z. Если при этом $kz\ll T$, то можно получить k ветвей разнесенного приема с примерно

---

[1] Очевидно, такие же соотношения будут иметь место, если задержать на z не сигнал, а напряжение ЛЧМ гетеродина.

одинаковым отношением сигнал/шум h^2 и с некоррелированными шумами (рис. 10.10). После сложения результатов детектирования получим отношение сигнал/шум около kh^2. При достаточно большом значении $\gamma=\Delta f/T$ можно обеспечить сколь

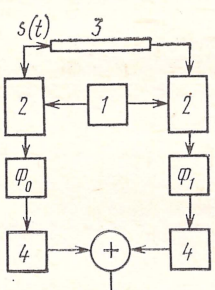


Рис. 10.9. Двухканальный прием ЛЧМ сигнала:
1 — ЛЧМ гетеродин; *2* — перемножитель; *3* — линия задержки; *4* — детектор; *Ф* — согласованные фильтры

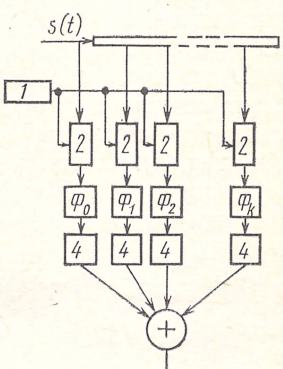


Рис. 10.10. Многоканальный («сверхоптимальный») прием ЛЧМ сигнала (обозначения те же, что и на рис. 10.9)

угодно большое k и, следовательно, сколь угодно большую помехоустойчивость приема по схеме рис. 10.10, которую тов. Маляров назвал схемой сверхоптимального приема.

На различных обсуждениях предложения Малярова чаще всего можно было услышать, что шумы в сформированных ветвях разнесения не будут некоррелированными. На первых порах тов. Маляров просто ссылался на общеизвестный факт, что процессы, выделенные из белого шума фильтрами

с непересекающимися полосами, некоррелированны. Когда же его внимание обратили на то, что в рассматриваемой схеме шум не просто разделяется фильтрами, а проходит сначала через перемножители и при этом становится нестационарным, он представил расчет функции взаимной корреляции шумов. Приведем этот расчет, ограничившись для простоты случаем двух ветвей (см. рис. 10.9).

Пусть шум $n(t)$ на входе является белым и имеет спектральную плотность N_0. После перемножителей получим напряжение $n(t)u_г(t)$ и $n(t-z)u_г(t)$. Обозначив импульсные характеристики фильтров Φ_0 и Φ_1 соответственно $g_0(t)$ и $g_1(t)$, получим напряжения шумов на выходах фильтров:

$$\left.\begin{aligned}\eta_0(t) &= \int_0^T g_0(x)\, n(t-x)\, u_г(t-x)\, dx,\\ \eta_1(t) &= \int_0^T g_1(y)\, n(t-z-y)\, u_г(t-y)\, dy.\end{aligned}\right\} \qquad (10.11)$$

Взаимная корреляционная функция этих процессов в совпадающие моменты времени равна по определению

$$R_{01}(0,\ t) = \langle \eta_0(t)\, \eta_1(t) \rangle = \int_0^T \int_0^T g_0(x)\, g_1(y) \langle n(t-x) \times n(t-z-y) \rangle u_г(t-x)\, u_г(t-y)\, dx dy. \qquad (10.12)$$

Здесь угловые скобки означают математическое ожидание.

Эта функция зависит не только от сдвига z, но и от t, так как процесс после перемножения нестационарный. Найдем среднее значение по времени

$$\widetilde{R_{01}(0,\ t)} = \frac{1}{T}\int_0^T dt \int_0^T\int_0^T g_0(x)\, g_1(y)\, u_\Gamma(t-x)\, u_\Gamma(t-$$

$$-y)\,\langle n(t-x)\, n(t-z-y)\rangle\, dxdy. \qquad (10.13)$$

Так как шум на входе белый и стационарный, то

$$\langle n(t-x)\, n(t-z-y)\rangle = 0{,}5N_0\,\delta(x-y-z). \qquad (10.14)$$

Подставляя (10.14) и (10.9) в (10.13) и интегрируя по t, получим [1]

$$\widetilde{R_{01}(0,\ t)} = R_{01}(0) = \frac{N_0}{2T}\int_0^T\int_0^T\int_0^T g_0(x)\, g_1(y)\,\delta(x-z-$$

$$-y)\, U_\Gamma^2 \cos[\omega_\Gamma(t-x) + \pi\gamma(t-x)^2 + \psi]\cos[\omega_\Gamma(t-y)+$$

$$+\pi\gamma(t-y)^2+\psi]\, dxdydt = \frac{N_0U_\Gamma^2}{4}\int_0^T\int_0^T \frac{\sin\pi\gamma(x-y)\,T}{\pi\gamma(x-y)\,T}\times$$

$$\times\, g_0(x)\, g_1(y)\,\delta(x-z-y)\cos[\omega_\Gamma(x-y) - \pi\gamma(x^2-y^2)+$$

$$+2\pi\gamma(x-y)\,T]\, dxdy;$$

далее, используя фильтрующее свойство δ-функции, произведем интегрирование по y

$$R_{01}(0) = \frac{N_0U_\Gamma}{4}\,\frac{\sin\pi\gamma zT}{\pi\gamma zT}\int_0^T g_0(x)\, g_1(x-z)\cos[\omega_\Gamma z -$$

$$-2\pi\gamma zx - \pi\gamma z^2 + 2\pi\gamma zT]\, dx. \qquad (10.15)$$

---

[1] При этом учтено, что интеграл от быстро осциллирующей функции с множителем $\cos\omega_\Gamma(2t-x-y)$ практически равен нулю.

Для согласованных фильтров

$$g_0(t) = \begin{cases} \cos \Delta\omega_0 t, & 0 < t \leqslant T, \\ 0\ , & t < 0;\ t > T; \end{cases}$$

$$g_1(t) = \begin{cases} \cos \Delta\omega_1 t, & 0 < t \leqslant T, \\ 0\ , & t < 0;\ t > T. \end{cases} \qquad (10.16)$$

Подставляя эти значения в (10.15), получим

$$R_{01}(0) = \frac{N_0 U_г^2}{4} \frac{\sin \pi\gamma zT}{\pi\gamma zT} \int_0^T \cos \Delta\omega_0 x \cos \Delta\omega_1 (x - z) \times$$

$$\times \cos (2\pi\gamma zx + \psi_1)\, dx, \qquad (10.17)$$

где через ψ_1 обозначена не зависящая от x величина:

$$\psi_1 = \omega_г z - \pi\gamma z^2 + 2\pi\gamma zT.$$

Произведя интегрирование и учитывая, что $\Delta\omega_0 - \Delta\omega_1 = 2\pi\gamma z$, а $(\Delta\omega_0 + \Delta\omega_1) T \gg 2\pi$, имеем

$$R_{01}(0) = \frac{N_0 U_г^2 T}{8} \frac{\sin \pi\gamma zT}{\pi\gamma zT} \cos (\Delta\omega_1 z - \psi_1). \qquad (10.18)$$

Пусть теперь отводы в линии задержки сделаны через интервалы времени $1/\Delta f$. Тогда для любой пары отводов $z = r/\Delta f$, где r — целое число (не равное нулю). Так как $\gamma = \Delta f/T$, то $\gamma zT = r$, откуда

$$R_{01}(0) = \frac{N_0 U_г^2 T}{8} \frac{\sin r\pi}{\pi} \cos (\Delta\omega_1 z - \psi) = 0. \qquad (10.19)$$

Итак, помехи на выходе фильтров действительно некоррелированны. Что же касается огибающих помех, то, используя известное выражение для коэффициента корреляции $\rho^*(\tau)$ огибающей через коэффициент корреляции процесса $\rho(\tau)$

$$\rho^*(\tau) = 0{,}92\rho^2(\tau) + O(\rho^4(\tau)), \qquad (10.20)$$

и замечая, что $\rho(0)=0$, согласно (10.17) получим $\rho^*=0$.

Таким образом, сформированные «ветви разнесения» имеют некоррелированные помехи и, увеличивая их число, можно сколь угодно повышать помехоустойчивость «сверхоптимального» приема.

10.5. Давайте разберемся

Надеюсь, читатель отчетливо понимает, что во всех трех описанных выше рассуждениях содержатся ошибки. Это не подлежит сомнению, поскольку совершенно строго установлено, что при белом гауссовском шуме существует предельная вероятность ошибки, которую при заданном множестве сигналов нельзя уменьшить никакими схемными ухищрениями и она во всех трех случаях выше той, которую берутся обеспечить «изобретатели». Тем не менее было бы неправильным ограничить их критику классической формулой: «Этого не может быть, потому что этого не может быть никогда», которой в сущности довольно часто пользуются многие выступающие при обсуждениях подобных предложений. Необходимо в каждом конкретном случае указать, в чем именно заключается ошибка.

Возможно, что читатель уже эти ошибки нашел, в таком случае он может этот раздел пропустить.

Итак, начнем с идеи Пивоваровой. Ее ошибка заключается в том, что она рассматривает среднюю мощность шума на выходе частотного детектора, не обращая внимания на ее распределение. Как известно, шум на выходе частотного детектора состоит из двух составляющих — нормальной и аномальной. Именно нормальная составляющая,

имеющая в первом приближении гауссовское распределение, «подавляется» сигналом, в результате чего при большом индексе модуляции происходит выигрыш в отношении сигнал/шум. Это подавление вызвано тем, что мгновенная частота суммы сигнала и помехи мало отличается от мгновенной частоты сигнала, если огибающая сигнала больше огибающей помехи. В этих условиях существует только нормальная составляющая шума на выходе детектора. Поэтому, увеличивая девиацию, можно уменьшить среднюю квадратичную погрешность воспроизведения аналогового (например, телефонного) сигнала, т. е. повысить помехоустойчивость.

Вторая аномальная составляющая имеет импульсный характер, импульсы возникают в те моменты, когда огибающая помехи на входе превосходит огибающую сигнала. В этом случае мгновенная частота суммы сигнала и помехи мало отличается от мгновенной частоты помехи. Это поясняет рис. 6.1, где показано, что при значительной разнице амплитуд двух синусоид (кривая *а*) мгновенная частота близка к частоте составляющей с большей амплитудой.

При приеме телефоного сигнала с ЧМ аномальные импульсы ощущаются как небольшие щелчки. Если отношение сигнал/помеха на входе детектора превышает некоторый пороговый уровень (зависящий от требуемого качества приема), то этими щелчками можно пренебречь, так как они возникают крайне редко. Ниже порогового уровня щелчки происходят часто и мощность аномальной составляющей помехи может значительно превзойти мощность нормальной составляющей.

Рассмотрим теперь прием дискретных ЧМ сигналов и предположим, что для сравнения один и

тот же принятый сигнал подан на две решающие схемы — с разделительными фильтрами (рис. 10.1) и с частотным детектором — ЧД (рис. 10.2). Пусть в течение некоторого тактового интервала T огибающая сигнала на входе решающих схем превышает огибающую помехи. В этом случае аномальные импульсы не возникают и на выходе ЧД будет только нормальный шум, причем отношение сигнал/шум на выходе детектора может быть в десятки и даже сотни раз (при большой девиации) больше, чем на входе. Конечно, в этих условиях схема ЧД обеспечивает безошибочный прием. Но при этом и в схеме с разделительными фильтрами ошибка невозможна, так как даже если вся помеха попадет в фильтр, в котором нет сигнала, то все же напряжение, созданное на фильтре с сигналом, будет больше напряжения на фильтре с помехой.

Ошибка в схеме с разделительными фильтрами может произойти с некоторой вероятностью (меньшей 0,5) только в тех тактовых интервалах, когда вследствие выброса шума его огибающая превзойдет огибающую сигнала. Очевидно, что в этом тактовом интервале на выходе ЧД возникнет аномальная помеха, так как мгновенная частота будет определяться шумом. При этом знак напряжения на выходе ЧД с почти равной вероятностью может быть положительным или отрицательным, так что в половине таких случаев произойдет ошибка.

Таким образом, «выигрыш» в отношении сигнал/шум в ЧД имеет место только на тех тактовых интервалах, когда ошибка не может произойти и в схеме с разделительными фильтрами. В тех же случаях, когда ошибка может произойти, ЧД дает не «выигрыш», а «проигрыш» и вероятность

ошибки в схеме рис. 10.2 будет не меньше, а, вообще говоря, больше, чем в схеме рис. 10.1, или в лучшем случае такая же. Увеличение индекса модуляции k не влияет на вероятность ошибок в схеме с разделительными фильтрами, а в схеме с ЧД не уменьшает, а увеличивает вероятность ошибок, так как в более широкой полосе частот больше мощность шума и чаще возникают аномальные импульсы.

Итак, только при передаче аналогового сообщения можно повысить помехоустойчивость (точность воспроизведения сообщения), увеличивая девиацию. В дискретных же системах, когда помехоустойчивость характеризуется вероятностью ошибки, такого эффекта не возникает.

Очень похожая ошибка вкралась и в рассуждения Дьяковского. Он учитывал только искажения, возникающие на выходе приемника сигналов с интервальной импульсной модуляцией и вызываемые сдвигом фронтов импульсов при сложении сигнала и шума. Но, помимо этого, ошибки могут возникать и в те моменты времени, когда импульс сигнала отсутствует, если выброс шума достигнет такой величины, что его огибающая пересечет порог срабатывания, как показано на рис. 10.11.

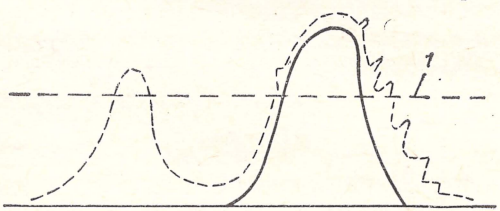


Рис. 10.11. Возникновение аномальных ошибок при интервальной манипуляции (*1* — пороговый уровень)

Кроме этого, могут возникать еще дополнительные ошибки в случаях, когда шум оказывается в противофазе с сигналом и огибающая импульса уменьшается настолько, что не пересекает порога срабатывания. Однако вероятность пропуска импульса обычно значительно меньше вероятности ложного срабатывания и поэтому будем учитывать только последнюю.

Заметим, что при ложном срабатывании в системе Дьяковского не просто искажается передаваемая буква, но и возникают две ложные буквы вместо одной переданной.

Оценим вероятность ложного срабатывания. Для гауссовского шума с дисперсией σ^2 одномерное распределение огибающей $X(t)$ является рэлеевским $w(X) = (X/\sigma^2)\exp(X^2/2\sigma^2)$.

При амплитуде импульса сигнала U_c порог срабатывания обычно устанавливается на уровне около $U_c/2$. Вероятность того, что огибающая шума в некоторый момент времени окажется выше этого порога, т. е. вероятность ложного срабатывания, равна

$$P_{л.с} = \int_{U_c/2}^{\infty} w(X)\,dX = \int_{U_c/2}^{\infty} (X/\sigma^2)\,e^{-(X^2/2\sigma^2)}\,dX =$$

$$= e^{-U_c^2/8\sigma^2} = e^{-h^2/4}, \qquad (10.21)$$

где последнее равенство написано в предположении, что мощность сигнала равна $U^2_c/2$, а отношение мощностей сигнал/помеха после согласованного фильтра равно h^2.

Белый шум, прошедший через согласованный фильтр, имеет интервал корреляции, равный T. Поэтому, учитывая, что для гауссовских процессов

некоррелированность означает независимость, значения огибающей, взятые через интервалы, кратные T, независимы. Следовательно, полагая, что между двумя импульсами сигнала интервал равен aT, вероятность появления хотя бы одного ложного срабатывания

$$P_{л.с}(aT) \geqslant 1-(1-e^{-h^2/4})^{[a]}, \qquad (10.22)$$

где $[a]$ — целая часть a.

Здесь знак неравенства определяется двумя соображениями. Во-первых, величина a может быть дробной и тогда $[a]<a$, во-вторых, если во всех выбранных точках отсчета огибающая шума не превышает порога срабатывания, то все же она может пересечь этот порог где-то между точками отсчета. Отсюда вероятность «аномальной» ошибки при некоторой средней длительности интервала

$$P_{ан} \geqslant 1-(1-e^{-h^2/4})^{\bar{a}}, \qquad (10.23)$$

где $\bar{a}$ — среднее значение a, приблизительно равное в нашем примере 6. Для рассмотренного примера $h=2$ получим $P_{ан}>0{,}96$. Даже при $h^2=16$ из этой формулы $P_{ан}>0{,}11$. В этих же условиях для двоичной системы ФМ при когерентном приеме из (10.6) имеем $P_{ош}\approx 2\cdot 10^{-7}$. Комментарии излишни.

Если сравнение двух систем производить не по пиковой, а по средней мощности, то для двоичной системы $h^2=16/5=3{,}2$ и тогда по ф-ле (10.6) $P_{ош}\approx 0{,}03$, т. е. все же меньше, чем в системе Дьяковского.

Конечно, повышение основания кода при прочих равных условиях позволяет повысить верность приема. Это отмечалось уже в [23]. Но сигналы и метод приема в системе Дьяковского выбраны далеко не лучшим образом, поэтому по сравнению с двоичной системой она не уменьшает, а увеличи-

вает вероятность ошибки. Такого рода сигналы с большой скважностью целесообразно использовать только в таких каналах, в которых в паузе отсутствует шум. Действительно, в квантовых каналах при отсутствии шумового фона и при очень низкой интенсивности сигнала их применение целесообразно.

Переходя к «сверхоптимальному приему» Малярова, заметим прежде всего, что в его доказательстве некоррелированности шумов на выходах фильтров различных отводов содержится «избыточность». Дело в том, что если подставить сразу значение $z=r/\Delta f$ в (10.15), а не в (10.18), как было сделано в его выводе, и учесть, что интеграл в (12.10) для любых импульсных реакций $g_0(t)$ и $g_1(t)$ физически реализуемых фильтров конечен [1], то окажется, что $R_{01}(0)=0$. Это значит, что шумы на выходах фильтров различных отводов в среднем некоррелированны, каковы бы ни были передаточные функции этих фильтров. Это уже наводит на размышления. Не зря еще в древности говорили: кто доказывает больше, чем нужно, тот ничего не доказывает.

Продолжая размышлять в этом направлении, рассмотрим два узкополосных стационарных гауссовских процесса с нулевым математическим ожиданием:

$$\left.\begin{aligned}\xi_1 &= A(t)\cos[\omega_0 t+\theta(t)],\\ \xi_2 &= A(t)\cos[(\omega_0+\nu)t+\theta(t)],\end{aligned}\right\} \qquad (10.24)$$

где $A(t)$ и $\theta(t)$ — низкочастотные стационарные случайные процессы, так что $\xi_1(t)$ отличается от $\xi_2(t)$ только сдвигом на частоту ν.

[1] Это следует из конечности области интегрирования и конечности модуля подынтегральной функции.

Функция взаимной корреляции в совпадающие моменты времени для ξ_1 и ξ_2 равна по определению

$$R_{1,2}(0,\ t) = \langle \xi_1(t)\,\xi_2(t) \rangle = 0{,}5 \langle A^2(t) \cos [(2\omega_0 + \nu)\,t + 2\theta(t)] \rangle + 0{,}5 \langle A^2(t) \cos \nu t \rangle. \qquad (10.25)$$

Эта функция зависит от времени t, хотя оба процесса стационарны. Дело в том, что они нестационарно связаны. Если теперь усреднить эту функцию по времени на бесконечном интервале, то, как легко убедиться, $\overline{R_{12}\ (0,\ t)} = 0$. Тот же результат получится и при усреднении на конечном интервале $T = 2n\pi/\nu$, если ω_0 кратна ν, или приближенно, если $\omega_0 \gg \nu$.

Можно ли, однако, из этого сделать вывод, что и огибающие этих процессов, хотя бы в среднем, некоррелированны? Ни в коем случае. Функция корреляции для огибающих в совпадающие моменты времени равна $\langle A^2(t) \rangle = 2\sigma^2$, где σ^2 — дисперсия процессов $\xi_1(t)$, $\xi_2(t)$, так что нормированная функция (коэффициент) корреляции $\rho^*(0) = 1$. Что же касается ф-лы (10.20), то она верна только для стационарно связанных гауссовских процессов.

Легко понять, что такой же результат получится и в том случае, если процесс $\xi_2(t)$ сдвинуть по времени на величину z, т. е. положить

$$\xi_2(t) = A(t - z) \cos [(\omega_0 + \nu)(t - z) + \theta(t - z)], \qquad (10.26)$$

с той лишь разницей, что теперь не $\rho^*(0)$, а $\rho^*(z) = 1$.

Заметим, что $\rho^*(0)$ очень мало отличается от $\rho^*(z)$, так как интервал корреляции огибающей порядка T, а по условию $z \ll T$.

Теперь все становится ясным. Огибающие шумов на выходах фильтров отводов в схеме рис. 10.10 жестко коррелированны. В этом проще всего можно убедиться, заменив схему рис. 10.9 совершенно эквивалентной по конечному результату схемой рис. 10.12. Здесь напряжение гетеродина, так же как и сигнал, задерживается на интервал времени z, так что на выходах перемножителей присутствуют совершенно одинаковые сигналы и шумы, но только сдвинутые на время z. Затем с помощью дополнительного преобразователя частоты сигнал и шум на выходе второго перемножителя сдвигаются по частоте на величину $\nu = 2\pi\gamma z$. После прохождения через фильтры Φ_1 и Φ_0 получатся в точности такие же напряжения, как и в точках на рис. 10.9.

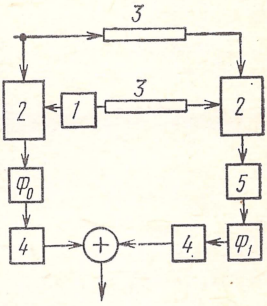


Рис. 10.12. Схема, позволяющая получить такие же выходные сигналы, как и схема рис. 10.9

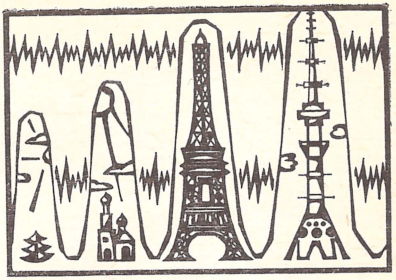
Но, с другой стороны, если шумы на выходе фильтра Φ_0 на рис. 10.12 обозначить как $\xi_1(t)$ в (10.24), то на выходе Φ_1 они представятся ф-лой (10.26) и, как было показано, их огибающие жестко коррелированны. Поэтому, сколько отводов линии задержки ни делать, все равно получится лишь много копий одного и того же сигнала с той же реализацией помехи, только сдвинутых по частоте и по времени. При сложении огибающих как сигналы, так и помехи будут складываться когерентно и никакого улучшения отношения сигнал/помеха не получится. Не спасает и то обстоятельство, что вследствие сдвига по времени коэффициент

корреляции огибающих в совпадающие моменты времени не в точности равен единице и, следовательно, при сложении огибающих двух сигналов значение шума не в точности удваивается. Это с избытком компенсируется тем обстоятельством, что и значение сигнала не удваивается, поскольку используемая энергия сигнала, сдвинутого на время z уменьшается на долю z/T.

Но может быть удастся воспользоваться тем, что жестко коррелированны только огибающие шума, а сами шумы в среднем некоррелированны? Что будет, если отказаться от некогерентного сложения, т. е. от амплитудных детекторов, и каким-либо образом осуществить когерентное сложение? Но и здесь ничего не получится. Непосредственно складывать сигналы, отличающиеся сдвигом по частоте, бессмысленно, так как их разность фаз будет непрерывно изменяться и когерентного сложения не получится. Если же осуществить дополнительное преобразование частоты и свести все сигналы к одинаковой частоте и фазе, то и помехи во всех отводах окажутся совершенно одинаковыми и когерентное сложение не даст выигрыша.

Итак, природу обмануть не удается. Минимальная достижимая вероятность ошибки при двоичной системе в канале с аддитивным белым гауссовским шумом однозначно определяется отношением энергии сигнала к спектральной плотности шума, и уменьшить ее нельзя.

Это же верно и для любой m-ичной системы, если $m>2$ — фиксированное число. Лишь в том случае, когда m ничем не ограничено, можно в соответствии с теорией Шеннона обеспечить сколько угодно малую вероятность ошибочного приема сообщения, если, конечно, скорость передачи меньше пропускной способности канала.



11

ПАРАДОКС ИМПУЛЬСНОЙ МОДУЛЯЦИИ

11.1. Порог помехоустойчивости

Начнем с хорошо известных фактов. При передаче непрерывных сообщений нелинейные методы модуляции позволяют получить «выигрыш» в отношении сигнал/шум. Это значит, что отношение мощности полезного сообщения на выходе демодулятора к спектральной плотности выходного шума (которым является погрешность оценки переданного сообщения) может быть больше отношения мощности сигнала к спектральной плотности помехи на входе приемника (или на выходе канала). Для простоты будем говорить только о канале с аддитивным гауссовским белым шумом [15]. Так, при ЧМ этот выигрыш равен $g'_{\text{ЧМ}}=3m^2/\Pi^2$, где m — индекс ЧМ, Π — пикфактор передаваемого сообщения, или, учитывая, что при больших индексах модуляции ширина спектра сигнала в канале $F_к$ приблизительно равна $2mF_c$, где F_c — ширина спектра сообщения,

$$g'_{\text{ЧМ}} \approx (3/4\Pi^2)(F_к/F_c)^2, \qquad (11.1)$$

При временно́й импульсной модуляции (ВИМ)

и оптимальном выборе формы импульса

$$g'_{\text{вим}} \approx (0{,}6/\Pi^2)(F_к/F_с)^2. \qquad (11.2)$$

Ряд других примеров можно найти в любом учебнике, например в [19]. Из них видно, что для увеличения выигрыша нужно расширять спектр сигнала.

Хорошо известно также, что обязательным условием получения этого выигрыша является достаточно высокая мощность сигнала $P_с$ на входе приемника. Существует пороговое значение мощности $P_{\text{пор}}$, такое, что при $P_с > P_{\text{пор}}$ можно пользоваться ф-лами (11.1), (11.2) и аналогичными им. При уменьшении $P_с$ ниже порогового выигрыш уменьшается и может даже стать меньше единицы. Так, при ЧМ получение выигрыша (11.1) определяется условием, что огибающая шума на входе ЧД не должна превышать огибающую сигнала. Конечно, если шум гауссовский, то его огибающая может достигать, в принципе, любых значений. Однако если вероятность того, что огибающая шума превысит амплитуду ЧМ сигнала A, меньше, скажем, 0,001, то с достаточным приближением можно пользоваться ф-лой (11.1).

Плотность распределения вероятностей огибающей B гауссовского шума с мощностью (дисперсией) $P_ш$, как известно, равна

$$w(B) = (B/P_ш)\exp(-B^2/2P_ш), \qquad (11.3)$$

и, следовательно, вероятность того, что $B > A$,

$$p\{B > A\} = \int_A^\infty w(B)\,dB = \exp(-A^2/2P_ш) =$$

$$= \exp(-P_с/P_ш),$$

где $P_с = A^2/2$ — мощность сигнала.

Положив эту вероятность равной 0,001, получим пороговое значение мощности сигнала из уравнения $\exp(-P_{пор}/P_ш) = 0{,}001$:

$$P_{пор} \approx 6{,}9P_ш. \qquad (11.4)$$

Если линейную часть приемника до входа частотного детектора (ЧД) можно считать полосовым фильтром с полосой пропускания, равной ширине спектра сигнала [1] $F_к$, то $P_ш = N_0F_к$ и

$$P_{пор} \approx 6{,}9N_0F_к. \qquad (11.5)$$

Таким образом, пороговая мощность ЧМ сигнала пропорциональна ширине его спектра. Если для увеличения выигрыша повышают индекс модуляции, что ведет к расширению спектра сигнала, то рано или поздно мощность сигнала окажется ниже пороговой и для дальнейшего улучшения верности приема придется ее увеличивать.

Иначе обстоит дело при некоторых импульсных методах модуляции, в частности при ВИМ. Момент прихода импульса регистрируется при пересечении его огибающей уровня срабатывания, который обычно устанавливается примерно на половине пикового значения импульса. Выигрыш (11.2) имеет место в том случае, когда в паузах между импульсами шум не достигает уровня срабатывания. При этом условии шум на выходе демодулятора возникает только в результате того, что аддитивная помеха перемещает положение фронта импульса («нормальный шум»). В противном случае возникает дополнительный «аномальный шум»,

[1] Здесь рассматривается «классическая» схема приема ЧМ сигналов. В современных помехоустойчивых системах используются следящие фильтры или следящие гетеродины, позволяющие сократить полосу пропускания и тем самым снизить $P_{пор}$.

имеющий импульсный характер и возникающий в результате того, что демодулятор воспринимает выброс помехи как импульс сигнала.

Пусть при некоторой ширине спектра сигнала и некотором уровне срабатывания помеха, прошедшая через входной фильтр приемника, имеет такую мощность $P_{ш}=N_0F_к$, что аномальный шум практически отсутствует. Предположим, что для увеличения выигрыша ширина спектра сигнала $F_к$ увеличена в n раз. Как видно из (11.2), выигрыш при этом должен возрасти в n^2 раз. Мощность помехи на входе демодулятора в результате расширения полосы пропускания возрастет в n раз, а ее выбросы — в $n^{1/2}$ раз. Может показаться, что теперь возникнет аномальный шум, так как выбросы помехи окажутся выше уровня срабатывания. Однако необходимо учесть, что при расширении спектра сигнала длительность $\tau_и$ импульса становится короче также в n раз, возрастает скважность и при сохранении средней мощности сигнала можно увеличить его пиковые значения в $n^{1/2}$ раз. Вследствие этого оптимальный уровень срабатывания повысится также в $n^{1/2}$ раз, т. е. пропорционально выбросам помехи. Следовательно, вероятность того, что величина шума превысит этот уровень, сохранится прежней. Другими словами, *если при ширине спектра $F_к$ аномальный шум практически отсутствовал, то он не возникнет и при ширине спектра $nF_к$.*

Вот что написано об этом, например, в [19]: «*Независимость* $P_{пор}$ от полосы частот — существенное преимущество импульсной модуляции по сравнению с ЧМ или ФМ. Выбирая, например, достаточно малое значение $\tau_и$ (т. е. достаточно широкополосный сигнал), можно увеличить выигрыш системы *без повышения* пороговой мощности сигна-

ла» (курсив мой — Л. Ф.). Такой же вывод можно найти в любом учебнике, где этот вопрос рассматривается. То же самое многократно говорил, да и продолжает говорить, автор в своих лекциях.

11.2. Пропускная способность

Посмотрим теперь, какие из этого можно сделать выводы. Предположим, что ВИМ используется для передачи непрерывных сообщений в канале с аддитивным белым гауссовским шумом, причем на ширину спектра сигнала не наложено никаких ограничений, а мощность P_c сигнала на выходе канала превышает пороговое значение $P_{пор}$, которое, как мы видели, не зависит от ширины спектра сигнала.

Зададимся сколь угодно малой величиной $\varepsilon > 0$ и потребуем, чтобы средний квадрат ошибки оценки сигнала (мощность «шума на выходе приемника») не превышал ε^2. При этом, как обычно, пиковое (или квазипиковое [1]) значение сообщения принято за единицу. Легко видеть, что это требование будет выполнено, если $g'_{ВИМ} \geqslant N_0 F_c / (P_c \varepsilon^2)$. Для этого выберем столь малую длительность импульса (или столь широкий спектр сигнала), чтобы в соответствии с (11.2) выполнилось неравенство

$$(0{,}6/\Pi^2)(F_к/F_c)^2 \geqslant N_0 F_c/(P_c \varepsilon^2),$$

или

$$F_к/F_c \geqslant 1{,}3\,(\Pi/\varepsilon)\sqrt{N_0 F_c/P_c}, \qquad (11.6)$$

что, в принципе, всегда можно сделать.

---

[1] Квазипиковым значением случайного процесса $x(t)$ называется величина $x_{кп}$, такая, что вероятность неравенства $|x(t)| > x_{кп}$ равна 0,001 или другой обусловленной величине.

При этом количество информации, передаваемое по каналу в единицу времени, равно эпсилон-энтропии сообщения H_ε при среднем квадрате ошибки ε^2, которая с уменьшением ε становится сколь угодно большой [1]. Отсюда следует, что канал с аддитивным гауссовским шумом в отсутствие ограничений на используемую полосу частот, если мощность сигнала выше некоторой постоянной, имеет бесконечную пропускную способность!

Но ведь это противоречит известной формуле для пропускной способности такого канала

$$\lim_{F_к \to \infty} C = \lim_{F_к \to \infty} F_к \log (1 + P_с/N_0 F_к) = (P_с/N_0) \log e, \quad (11.7)$$

из которой видно, что при конечной мощности сигнала пропускная способность при любой ширине спектра остается конечной. Следовательно, в наших рассуждениях есть ошибки.

11.3. Небольшие уточнения

Действительно, при более тщательном проведении рассуждений удается обнаружить даже не одну, а две небольшие ошибки. Собственно говоря, единственный непреложный вывод из рассуждений о пороге помехоустойчивости состоит в том, что вероятность возникновения аномального импульса (т. е. пересечения огибающей помехи и уровня срабатывания) не зависит от ширины спектра сигнала. Оценим эту вероятность. Пусть A — максимальное значение огибающей импульса. Тогда

---

[1] Если, например, сообщение является гауссовским процессом с равномерным в полосе $F_с$ спектром, то $H_\varepsilon = F_с \log (1/\varepsilon^2) = -2F_с \log \varepsilon$.

$A/2$ — уровень срабатывания, и согласно (11.3)

$$p(B > A/2) = \int_{A/2}^{\infty} w(B)\,dB = \exp\{-A^2/(8P_ш)\}. \qquad (11.8)$$

Мощность сигнала

$$P_с = k_1 (A^2/2)(\tau_и/T), \qquad (11.9)$$

где k_1 — коэффициент, близкий к 1, определяемый формой импульса и не зависящий от ширины его спектра; $\tau_и$ — длительность импульса, отсчитываемая, например, на уровне $A/2$; T — средний период следования импульсов сигнала [интервал Котельникова $1/(2F_с)$].

Из (11.8) и (11.9) имеем

$$p(B > A/2) = \exp\{-P_с T/(4k_1 \tau_и P_ш)\}.$$

Заменив $P_ш$ его значением $N_0 F_к$ и учитывая, что $F_к = k_2/\tau_и$, где k_2 — также коэффициент, близкий к 1, получим окончательно

$$p(B > A/2) = \exp\{-(T/4k_1k_2)(P_с/N_0)\}. \qquad (11.10)$$

Итак, вероятность появления аномального импульса определяется отношением энергии сигнала за интервал Котельникова $T = 1/2F_с$ к спектральной плотности шума и не изменяется при сокращении длительности импульса и соответствующего расширения его спектра.

Далее, однако, было сделано два молчаливых допущения, которые при ближайшем рассмотрении оказываются неверными:

а) мощность аномального шума в системе ВИМ, которой можно практически пренебречь (пороговая мощность), не зависит от отношения $F_к/F_с$;

б) мощность аномального шума зависит только от вероятности появления аномального импульса.

Ошибочность первого допущения сразу становится очевидной, если сравнить мощности аномального и нормального шумов. Мощность нормального шума $P_н=\varepsilon^2$ и согласно (11.6) может быть сколь угодно уменьшена при увеличении отношения $F_к/F_с$. Если бы при этом мощность аномального шума $P_{ан}$ оставалась постоянной, то с расширением спектра сигнала мы рано или поздно пришли бы к положению, когда $P_н \ll P_{ан}$, и дальнейшее расширение спектра уже не могло бы улучшить верность приема сообщения. Поэтому пороговую мощность аномального шума нельзя устанавливать одинаковой для всех значений $P_н$. Если, например, при $P_н=10^{-4}$ и $P_{ан}=10^{-6}$ можно аномальным шумом полностью пренебречь, то при $P_н=10^{-7}$ и том же значении $P_{ан}$ именно аномальный шум является определяющим.

Разумно принять за условное определение порога помехоустойчивости такое отношение сигнал/шум, когда $P_{ан}=P_н$, т. е. выигрыш оказывается вдвое меньше расчетного.

Второе допущение также ошибочно. Для определения мощности аномального шума недостаточно знать вероятность аномального импульса (11.10). Для того чтобы получить об этом хотя бы общее представление, рассмотрим случай, когда модуляции нет, т. е. сигнал представляет собой равномерную последовательность импульсов, возникающих в моменты времени nT $(n=0, 1, 2, \ldots)$ с длительностью $\tau_и \ll T$. Широко применяемый метод приема сигналов ВИМ заключается в том, что выборочное значение оценки $x^*(nT)$ принятого сообщения полагается равным $(t_n-nT)/T$, где t_n — момент первого пересечения импульсом уровня срабатывания огибающей приходящего сигнала на интервале $(n-1/2)T<t<(n+1/2)T$. Именно для

такого метода приема вычислен выигрыш (11.2). Легко видеть, что в отсутствие помех такой метод позволяет точно восстановить переданное выборочное значение сообщения $x(nT)$. Аномальный шум возникает, если до прихода импульса сигнала, т. е. на интервале $(n-1/2)T<t<nT$, огибающая помехи пересечет уровень срабатывания.

При ширине спектра помехи $F_к$, полагая, что спектр ее в этой полосе равномерен, в каждую секунду существует $2F_к$ независимых отсчетов помехи. В интервале времени от $(n-0{,}5)T$ до nT существует $F_кT$ независимых отсчетов помехи. В первом приближении можно считать, что аномальный импульс возникает в том случае, если хотя бы один из этих отсчетов превысит уровень срабатывания $A/2$. Вероятность этого равна $1-(1-p)^{F_кT} \approx pF_кT$, где через p обозначена вероятность $p(B>A/2)$, определяемая ф-лой (11.10), а приближенное равенство написано в предположении $pF_кT \ll 1$, что всегда выполняется при допустимом качестве приема. Заметим, что при этом предположении можно совершенно не считаться с возможностью появления более одного аномального импульса на одном тактовом интервале.

Очевидно, что аномальный импульс может появиться с равной вероятностью в любой момент времени t на интервале от $(n-1/2)T$ до nT. При этом возникнет ошибка оценки сообщения $\Delta x = x^*-x=(t-nT)/T$. Таким образом, условная плотность вероятности момента прихода аномального импульса

$$w(t) = 2/T, \qquad (n-1/2)T<t<nT.$$

Средний квадрат этой ошибки (при условии, что она произошла)

$$<\Delta x^2> = \int\limits_{(n-1/2)\,T}^{nT} \left(\frac{t-nT}{T}\right)^2 w(t)\,dt =$$

$$= \int\limits_{(n-1/2)\,T}^{nT} \frac{2(t-nT)^2}{T^3}\,dt = \frac{1}{12}.$$

С учетом вероятности появления аномального импульса, вычисленной выше, средняя мощность аномального шума на выходе

$$P_{\text{ан}} \approx pF_{\text{к}}T<\Delta x^2> = \frac{F_{\text{к}}T}{12}\exp\left\{-\frac{T}{4k_1k_2}\,\frac{P_{\text{с}}}{N_0}\right\}. \quad (11.11)$$

Таким образом, мощность аномального шума определяется не одной вероятностью p, а еще и множителем, пропорциональным $F_{\text{к}}$.

Чтобы довести исследование до конца, оценим пороговое значение мощности сигнала. Согласно условию порог помехоустойчивости имеет место при $P_{\text{ан}}=\varepsilon^2$ или из (11.6) и (11.11) при

$$1{,}67\,\frac{\Pi^2 N_0 F_{\text{с}}^3}{P_{\text{с}}F_{\text{к}}^2} = \frac{F_{\text{к}}T}{12}\exp\left\{-\frac{T}{4k_1k_2}\,\frac{P_{\text{с}}}{N_0}\right\}. \quad (11.12)$$

Полагая $k_1=k_2=1$ и обозначая, как обычно, $P_{\text{пор}}T/N_0=h^2_{\text{пор}}$, $F_{\text{к}}/F_{\text{с}}=\alpha$, а также учитывая, что $T=1/2F_{\text{с}}$, сведем это уравнение к виду

$$h^2_{\text{пор}}/4 - \ln h^2_{\text{пор}} = \ln(\alpha^3/30{,}7\,\Pi^2). \quad (11.13)$$

При больших значениях α (например, при $\alpha>1000$) пороговое значение $h^2_{\text{пор}}$ много больше 1 и в левой части можно пренебречь $\ln h^2_{\text{пор}}$ по сравнению с $h^2_{\text{пор}}/4$. Тогда, возвращаясь к пороговой мощности сигнала $P_{\text{пор}}=h^2_{\text{пор}}N_0/T$, получим асимптотически

$$P_{\text{пор}} \approx N_0F_{\text{с}}\ln(\alpha^3/30{,}7\,\Pi^2). \quad (11.14)$$

Таким образом, пороговая мощность сигнала ВИМ с увеличением α не остается постоянной, а возрастает. Для получения сколь угодно большой верности передачи непрерывного сообщения недостаточно увеличивать α, как следовало из (11.6), в пренебрежении аномальным шумом, но необходимо также увеличивать P_c в соответствии со смыслом ф-лы (11.7), и «парадокс» полностью снимается.

Можно ли на основании этого утверждать, что системы с импульсной модуляцией не имеют преимуществ перед такими, как ФМ или ЧМ? Конечно, нельзя. Сравнивая (11.14) с (11.5), замечаем, что в системе ЧМ пороговая мощность сигнала растет пропорционально $P_ш = N_0 F_к$, т. е. пропорционально α, тогда как при ВИМ зависимость пороговой мощности от α при $\alpha \gg 1$ логарифмическая [4]. На рис. 11.1 показана зависимость $P_{пор}/N_0 F_c$ от α при $\Pi^2 = 10$ для ЧМ и ВИМ. Там же показаны зависимости g' от α для тех же систем. Пороговая мощность для ВИМ возрастает по срав-

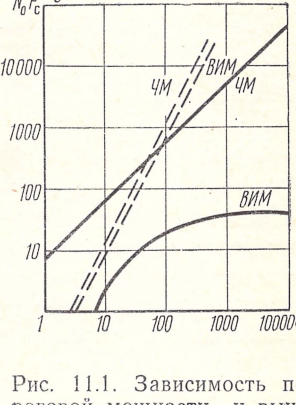


Рис. 11.1. Зависимость пороговой мощности и выигрыша от ширины спектра сигнала для ЧМ и ВИМ: _____ пороговая мощность; — — — — выигрыш

[1] Рекомендуем читателю доказать (пользуясь понятием «эпсилон-энтропия»), что из ф-лы (11.7) также следует логарифмическая зависимость необходимой мощности сигнала от средней квадратичной погрешности ε, которая при ВИМ пропорциональна $1/\alpha$.

нению с ЧМ настолько медленно, что ее можно действительно считать почти постоянной. Именно это позволяет использовать при ВИМ очень широкий спектр сигнала и осуществлять выигрыш g' порядка нескольких тысяч. При ЧМ это практически бессмысленно, так как пороговая мощность оказывается столь большой, что выгоднее применять узкополосные методы модуляции, например однополосную.

Таким образом, фраза, набранная курсивом на стр. 198, неверна. Мощность сигнала, при которой аномальный шум «практически отсутствует», несколько возрастает с расширением спектра. Тем не менее цитата из учебника, приведенная там же, в первом приближении справедлива. Можно было бы уточнить ее, заменив набранные курсивом слова: вместо «независимость» написать «очень слабая зависимость», а перед «без повышения» добавить слово «почти».

Практически при расчете системы связи с ВИМ это незначительное повышение порога помехоустойчивости можно вовсе не учитывать. Но для теории оно очень важно, так как только его учет согласует верность приема с пределом пропускной способности канала.

12

БЕРЕГИСЬ НЕТОЧНЫХ ФОРМУЛИРОВОК!

12.1. Занимательные парадоксы

Все мы еще в школе удивлялись различного рода «доказательствам» того, что $2 \times 2 = 5$ или все числа равны друг другу и т. п. Конечно, никто эти утверждения всерьез не принимает и цель таких доказательств заключается в том, чтобы научить юных математиков отыскивать логические или аналитические ошибки в доказательствах и подчеркнуть необходимость тщательных и точных формулировок. Чаще всего такие арифметические миниатюры основываются на неявном делении на нуль, на неучете двузначности операции извлечения квадратного корня, на умножении неравенства на отрицательное число без изменения знака неравенства, на некорректном применении полной математической индукции и т. п.

Здесь будут изложены два такого же рода «доказательства» нелепых результатов, относящихся к теории связи. В отличие от ошибок, описанных в других главах, никто никогда эти результаты не считал верными. Однако ошибки в рассуждениях здесь не сразу бросаются в глаза и отыскание их

может быть полезным для более глубокого понимания некоторых свойств гауссовских случайных величин и процессов.

12.2. Количество информации, передаваемой в гауссовском канале с неопределенной фазой

Рассмотрим непрерывный канал связи с неопределенной фазой, пропускающий сигналы в некоторой полосе частот F со средней круговой частотой ω_0. При прохождении сигнала через канал фаза всех его составляющих поворачивается на одинаковый случайный угол Δ, равномерно распределенный на интервале $(-\pi, \pi)$ и не зависящий от передаваемого сигнала. Аддитивной помехи нет (или ею можно пренебречь).

Будем подавать на вход канала узкополосный гауссовский процесс

$$x(t) = A(t)\cos[\omega_0 t + \theta(t)] = \alpha(t)\cos\omega_0 t - \beta(t) \times$$
$$\times \sin\omega_0 t = \mathrm{Re}\left\{[\alpha(t) + i\beta(t)]\, e^{i\omega_0 t}\right\} = \mathrm{Re}\left\{\xi(t)\, e^{i\omega_0 t}\right\}, \tag{12.1}$$

где $\alpha(t) = A(t)\cos\theta(t)$; $\beta(t) = A(t)\sin\theta(t)$ — независимые гауссовские процессы с нулевым математическим ожиданием и одинаковой дисперсией, которую обозначим σ^2; $\xi(t) = \alpha(t) + i\beta(t)$ — комплексный гауссовский процесс [комплексная огибающая процесса $x(t)$].

Сигнал на выходе канала

$$y(t) = A(t)\cos[\omega_0 t + \theta(t) + \Delta] = [\alpha(t)\cos\Delta - \beta(t) \times$$
$$\times \sin\Delta]\cos\omega_0 t - [\beta(t)\cos\Delta + \alpha(t)\sin\Delta]\sin\omega_0 t =$$
$$= \mathrm{Re}\left\{\eta(t)\, e^{i\omega_0 t}\right\}, \tag{12.2}$$

где

$$\eta(t) = \alpha(t)\cos\Delta - \beta(t)\sin\Delta + i[\beta(t)\cos\Delta + \alpha(t)\sin\Delta] = \xi(t)e^{i\Delta} \quad (12.3)$$

— также гауссовский комплексный процесс [комплексная огибающая $y(t)$].

Количество информации, содержащееся в некотором отсчете y относительно отсчета x, $I(y, x)$, очевидно совпадает с количеством информации $I(\eta, \xi)$, содержащимся в отсчете η относительно ξ, поскольку комплексные огибающие (при фиксированной частоте ω_0) полностью определяют соответствующие сигналы.

Найдем коорреляционный момент

$$K(\eta, \xi) = \langle\eta\xi^*\rangle = \langle|\xi|^2 e^{i\Delta}\rangle,$$

где угловые скобки означают усреднение по ансамблю.

Так как по условию ξ и Δ — независимые случайные величины, то

$$\langle|\xi|^2 e^{i\Delta}\rangle = \langle|\xi|^2\rangle\langle e^{i\Delta}\rangle = \langle|\xi|^2\rangle\langle\cos\Delta + i\sin\Delta\rangle = \langle|\xi|^2\rangle\left[\int_{-\pi}^{\pi}\frac{1}{2\pi}\cos\Delta\, d\Delta + i\int_{-\pi}^{\pi}\frac{1}{2\pi}\sin\Delta\, d\Delta\right] = 0. \quad (12.4)$$

Таким образом ξ и η некоррелированные, а поскольку они гауссовские, то и независимые случайные величины [1]. Так, в известной книге [26] на

[1] Легко убедиться, что и отсчеты $\xi(t_1)$ и $\eta(t_2)$, взятые в различные моменты времени, также некоррелированны, так как $\langle\xi^*(t_1)\eta(t_2)\rangle = \langle\xi^*(t_1)\xi(t_2)\rangle\langle e^{i\Delta}\rangle = 0$.

с. 64 сказано: «...для двух *нормальных* случайных величин *некоррелированность* означает также *независимость*» (курсив автора).

С другой стороны, хорошо известно (см., например, теорему 2.3.2 [10]), что для статистически независимых ξ и η $I(\eta, \xi)=0$, а отсюда следует, что $I(y, x)=0$. Итак, мы показали, что в канале с неопределенной фазой при подаче на его вход гауссовского сигнала информация не передается даже в отсутствие аддитивного шума. Не правда ли, странный результат?

В самом деле, если продетектировать сигнал $y(t)$, то можно точно восстановить огибающую $A(t)$ переданного сигнала $x(t)$ и, следовательно, получить сколь угодно большую информацию. Рассуждая более формально, можно исходить из известного и очевидного неравенства

$$I(y, x) \geqslant I(|\eta|, |\xi|) = I(A, A) = H(A) = \infty, \qquad (12.5)$$

так как A принимает значения на континууме. В реальных условиях вследствие неизбежного аддитивного шума эта информация конечна, но может быть сколь угодно велика при достаточно слабом шуме и уж во всяком случае не равна нулю при конечном отношении мощностей сигнал/шум.

Таким образом, мы получили два противоположных результата — взаимная информация $I(y, x)$ согласно первому выводу равна нулю, а согласно второму — сколько угодно велика. Всякому, кто передавал информацию по радио или другим высокочастотным каналам и использовал некогерентный прием, ясно, что ошибочен первый вывод. Но в чем же ошибка?

Многие из тех, кому автор задавал этот вопрос, начинали сомневаться в том, что отсчет η является гауссовской случайной величиной. Однако эти сом-

нения легко рассеиваются. Обозначим $\lambda = \alpha \cos\Delta - \beta \sin\Delta$; $\mu = \beta \cos\Delta + \alpha \sin\Delta$, где α, β и Δ — совместно независимы. Тогда плотность вероятности λ

$$w(\lambda) = \int_{-\pi}^{\pi} w(\lambda \mid \Delta)\, w(\Delta)\, d\Delta = (1/(2\pi)) \int_{-\pi}^{\pi} w(\lambda \mid \Delta)\, d\Delta.$$

Но $w(\lambda|\Delta)$ представляет собой плотность вероятностей разности $\alpha \cos\Delta - \beta \sin\Delta$ при некотором фиксированном Δ, каждый из членов является гауссовской случайной величиной с дисперсиями соответственно $\sigma^2 \cos^2\Delta$ и $\sigma^2 \sin^2\Delta$ и, следовательно, λ — также гауссовская величина с дисперсией σ^2. Аналогично то же самое можно показать и относительно μ. Кроме того, легко убедиться, что λ и μ статистически независимы. Таким образом, $\eta = \lambda + i\mu$ является комплексной гауссовской величиной, так что это возражение отпадает.

Суть дела проясняется, если рассмотреть четыре действительные гауссовские величины: α, β, λ и μ. Нетрудно видеть, что они попарно некоррелированны и независимы. Более того, любые три из этих величин имеют совместное нормальное распределение вероятностей, например:

$$w(\alpha,\ \beta,\ \lambda) = w(\alpha)\, w(\beta)\, w(\lambda) = \frac{1}{(2\pi\sigma^2)^{3/2}} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma^2}(\alpha^2 + \beta^2 + \lambda^2)\right\}. \qquad (12.6)$$

Но как только мы попробуем учесть еще и четвертую величину, то сразу наталкиваемся на зависимость

$$\alpha^2 + \beta^2 = \lambda^2 + \mu^2 = A^2. \qquad (12.7)$$

Таким образом, значение $|\mu|$ полностью определяется значениями α, β, λ. Поэтому совместную четы-

рехмерную плотность вероятностей для них следует записать так:

$$w(\alpha, \beta, \lambda, \mu) = \frac{1}{2(2\pi\sigma^2)^{3/2}} \exp\left\{-\frac{1}{2\sigma^2}(\alpha^2+\beta^2+\lambda^2)\right\} \times$$

$$\times [\delta(\mu - \sqrt{\alpha^2+\beta^2-\lambda^2}) + \delta(\mu + \sqrt{\alpha^2+\beta^2-\lambda^2})]. \quad (12.8)$$

«Лишняя» двойка в знаменателе учитывает равную вероятность значений $+\mu$ и $-\mu$.

Вспомним теперь, как доказывается эквивалентность понятий «некоррелированность» и «независимость» для нормальных случайных величин. Для этого записывается совместное нормальное распределение n величин, корреляционная матрица полагается диагональной (что означает попарную некоррелированность) и тогда оказывается, что n-мерная плотность распадается на произведение n одномерных плотностей (что означает независимость). Поэтому правильная формулировка этого свойства должна звучать так: «для *совместно нормальных* случайных величин некоррелированность означает также независимость». Слово «совместно» часто опускают, и в результате возникает неточность.

В нашем случае ξ и η, будучи каждая в отдельности гауссовской, совместно гауссовскими не являются, так как четырехмерное распределение для их составляющих (12.8) негауссовское. Поэтому, хотя ξ и η некоррелированны, они не являются независимыми и взаимная информация между ними не равна нулю.

12.3. О рэлеевских замираниях

Канал с рэлеевскими замираниями можно определить следующим образом. Если на вход канала по-

дан сигнал $\cos\omega_1 t$, то на выходе наблюдается сигнал

$$z_1(t) = \mu_{c1}\cos\omega_1 t + \mu_{s1}\sin\omega_1 t, \qquad (12.9)$$

где $\mu_{c1}(t)$ и $\mu_{s1}(t)$ — стационарные и стационарно связанные гауссовские процессы с нулевым математическим ожиданием и одинаковыми автокорреляционными функциями, вообще говоря, зависящие от ω_1. Кроме того, для них по определению имеет место соотношение $<\mu_{c1}(t)\,\mu_{s1}(t')> = -<\mu_{s1}(t)\times \times\mu_{c1}(t')>$, откуда следует также, что $<\mu_{c1}(t)\,\mu_{s1}(t)> = 0$, т. е. в одинаковые моменты времени μ_{c1} и μ_{s1} взаимно некоррелированны. Будем полагать μ_{c1} и μ_{s1} «медленными» процессами. Это значит, что их нормированные автокоррелированные функции

$$<\mu_{c1}(t)\,\mu_{c1}(t+\tau)>/<\mu_{c1}^2>$$

практически не отличаются от 1 для значений τ, меньших периода частоты входного сигнала $T_1 = 2\pi/\omega_1$.

Очевидно, что $z_1(t)$ представляет также гауссовский процесс (узкополосный). Менее очевидно, что процесс $z_1(t)$ стационарный в широком смысле. Однако в этом можно убедиться, вычислив автокорреляционную функцию $z_1(t)$ для двух моментов времени t и t':

$$B_{z1}(t,\ t') = <z_1(t)\,z_1(t')> = <[\mu_{c1}(t)\cos\omega_1 t + \mu_{s1}(t)\sin\omega_1 t[\times[\mu_{c1}(t')\cos\omega_1 t' + \mu_{s1}(t')\sin\omega_1 t']> =$$
$$= <\mu_{c1}(t)\,\mu_{c1}(t')\cos\omega_1 t\cos\omega_1 t' + \mu_{c1}(t)\,\mu_{s1}(t')\times$$
$$\times\cos\omega_1 t\sin\omega_1 t' + \mu_{s1}(t)\,\mu_{c1}(t')\sin\omega_1 t\cos\omega_1 t' +$$
$$+\mu_{s1}(t)\,\mu_{s1}(t')\sin\omega_1 t\sin\omega_1 t'>. \qquad (12.10)$$

Учитывая, что $<\mu_{c1}(t)\,\mu_{c1}(t')> = <\mu_{s1}(t)\,\mu_{s1}(t')> = B_\mu(t'-t)$, $-<\mu_{c1}(t)\,\mu_{s1}(t')> = <\mu_{s1}(t)\,\mu_{c1}(t')> = \widetilde{B}_\mu(t'-t)$, это выражение упрощается:

$$B_z(t,\ t') = B_\mu(t'-t)\cos\omega_1(t'-t) + \widetilde{B}_\mu(t'-t)\times$$
$$\times \sin\omega_1(t'-t). \qquad (12.11)$$

Таким образом, корреляционная функция $B_z(t, t')$ зависит только от разности $t'-t$, а так как математическое ожидание $z_1(t)$ равно нулю, т. е. постоянно, то $z_1(t)$ является стационарным в широком смысле гауссовским процессом.

Подадим теперь на вход канала сигнал, состоящий из двух гармонических составляющих: $\cos\omega_1 t + \cos\omega_2 t$, $\omega_1 \neq \omega_2$. На выходе сигнал представляет сумму двух стационарных гауссовских процессов:

$$z(t) = z_1(t) + z_2(t) = \mu_{c1}(t)\cos\omega_1 t + \mu_{s1}(t)\sin\omega_1 t +$$
$$+ \mu_{c2}(t)\cos\omega_2 t + \mu_{s2}(t)\sin\omega_2 t. \qquad (12.12)$$

Эта сумма является также гауссовским процессом.

Включим теперь на выходе канала два фильтра с амплитудно-частотными характеристиками, близкими к прямоугольным, и неперекрывающимися полосами пропускания так, чтобы один из них пропустил сигнал $z_1(t)$, а другой — $z_2(t)$. Это всегда можно сделать со сколь угодно большой точностью, даже при очень близких частотах ω_1 и ω_2, если все $\mu(t)$ изменяются очень медленно. Так, если спектр $\mu(t)$ занимает частоты ниже 1 Гц, как это часто бывает в коротковолновых радиоканалах, то частоты ω_1 и ω_2 могут различаться, скажем, на 10 Гц.

Для стационарных случайных процессов доказано (см., например, [41]), что сигналы, выделенные такими фильтрами с неперекрывающимися полосами пропускания, взаимно некоррелированны при любых сдвигах времени. Следовательно, можно утверждать, что $z_1(t)$ и $z_2(t)$ взаимно некоррелированны по крайней мере с той точностью, с ка-

кой их можно разделить фильтрами. Но если некоррелированны гауссовские процессы $z_1(t)$ и $z_2(t)$, то они и независимы, а следовательно, независимы и их огибающие $\mu_1 = \sqrt{\mu^2_{c1} + \mu^2_{s1}}$ и $\mu_2 = \sqrt{\mu^2_{c2} + \mu^2_{s2}}$, представляющие коэффициенты передачи канала для двух близких частот: ω_1 и ω_2. Следовательно, даже при очень малых сдвигах по частоте рэлеевские замирания должны быть практически полностью селективными!

Здесь тоже явно что-то не так. Инженеры, занимающиеся коротковолновой радиосвязью, знают, что для того, чтобы замирания на двух частотах были слабо коррелированными, разность частот должна достигать по крайней мере 200—300 Гц, а иногда и более. Почему же мы получили неверный результат?

Не будем долго интриговать читателя и раскроем секрет. Все дело в том, что, хотя $z_1(t)$ и $z_2(t)$ — стационарные гауссовские процессы, их сумма не является стационарным процессом. В этом легко убедиться, вычислив функцию автокорреляции для $z_1(t) + z_2(t)$, аналогично тому, как это сделано в (12.10). При этом получатся «перекрестные произведения», т. е. члены вида $\mu_{c1}(t)\mu_{c2}(t') \cos \omega_1 t \cos \omega_2 t'$ и аналогичные, которые никак нельзя представить функциями только разности $t' - t$.

Теорема о некоррелированности сигналов на выходах неперекрывающихся спектров к нестационарному процессу $z(t)$ неприменима. Лишь в том случае, когда $\mu_{c1}(t)$ и $\mu_{s1}(t)$ некоррелированны соответственно с $\mu_{c2}(t')$ и $\mu_{s2}(t')$, перекрестные произведения имеют математическое ожидание, равное нулю, и автокорреляционная функция для $z(t)$ будет зависеть только от $t' - t$. Но в этой ситуации, когда $z(t)$ будет стационарным процессом, замирания на частотах ω_1 и ω_2 действительно независимы

и ничего удивительного в полученном результате не будет.

К сожалению, ошибочное мнение о стационарности суммы двух стационарных процессов довольно распространено среди инженеров и на этой почве нередко возникают недоразумения.

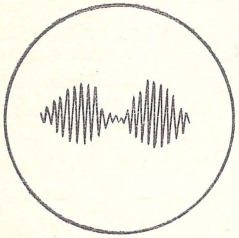


Рис. 12.1. Сумма двух реализаций стационарного шума, отличающихся сдвигом по частоте

Поучительным примером является поведение суммы центрированного стационарного процесса $x(t)$ с процессом $x_\Omega(t)$, полученным путем сдвига спектра исходного процесса на некоторую частоту Ω. Процесс, сдвинутый по частоте, можно представить в виде

$$x_\Omega(t) = x(t)\cos\Omega t + \hat{x}(t)\sin\Omega t, \qquad (12.13)$$

где $\hat{x}(t)$, как и ранее, — преобразование Гильберта от $x(t)$. Он, конечно, также является стационарным.

Суммарный процесс $z(t) = x(t) + x(t)\cos\Omega t + \hat{x}(t)\sin\Omega t$ нестационарен. Чтобы убедиться в этом, достаточно вычислить его дисперсию

$$\langle z^2(t)\rangle = 2D[1 + \cos\Omega t], \qquad (12.14)$$

где $D = \langle x^2(t)\rangle = \langle \hat{x}^2(t)\rangle$ — дисперсия исходного процесса.

Из (12.14) видно, что процесс $z(t)$ нестационарен, так как его дисперсия зависит от времени, изменяясь от нуля до $4D$. Это можно очень эффектно продемонстрировать на осциллографе. Сумма $x(t) + x_\Omega(t)$ имеет вид шума, модулированного по амплитуде по закону $\sqrt{1+\cos\Omega t} = |\sqrt{2}\cos(\Omega t/2)|$ (рис. 12.1).

13

КАК ОЦЕНИВАТЬ ВЕРНОСТЬ ПЕРЕДАЧИ СООБЩЕНИЙ?

13.1. Вероятность ошибочного декодирования

Вопрос, поставленный в заголовке главы, не так прост, как это может показаться. Когда-то о нем велись бурные споры, да и сейчас еще нередко приходится встречаться с необоснованными мнениями.

Математики, занимающиеся теорией кодирования, интересуются главным образом асимптотическим поведением системы, например, при увеличении длины кодового блока. Оценивая два метода кодирования, они сравнивают, насколько быстро уменьшается вероятность ошибочного декодирования $p_{\text{о.д}}$ при росте длины блока n. Как известно, в первом приближении вероятность ошибочного декодирования уменьшается с ростом n по экспоненциальному закону. Коэффициент E при n в показателе степени зависит от избыточности кода, измеряемой скоростью передачи R, которая в блочных кодах равна отношению k/n, где k — число информа-

ционных символов в блоке длиной n:

$$p_{\text{о.д}} = A\,\mathrm{e}^{-E(R)\,n}. \qquad (13.1)$$

Коэффициент A обычно также зависит от R и от n, впрочем довольно слабо. Поэтому решающую роль при сравнении двух методов кодирования играет величина $E(R)$. Чем она больше, тем лучше метод кодирования, по крайней мере, для больших n.

Против такого метода сравнения нельзя ничего возразить, если действительно рассматривается поведение кода при возрастании n. Однако нередко две системы сравниваются по величине $p_{\text{о.д}}$ при фиксированных значениях n, к тому же различных. Это иногда приводит к необоснованным суждениям.

Простейшим примером является такого рода рассуждение о целесообразности применения помехоустойчивого кода, с которым автору приходилось встречаться в некоторых статьях и авторефератах. Пусть в дискретном двоичном канале вероятность ошибки равна p. При кодировании блочным кодом длиной n вероятность ошибочного декодирования равна $p_{\text{о.д}}$ (впрочем, ее иногда обозначают так же, как и вероятность ошибки в канале, усугубляя этим путаницу). Тогда, рассуждают авторы этих статей, условием целесообразности кодирования является неравенство $p_{\text{о.д}} < p$ и чем оно сильнее выполняется, тем лучше код. Если же $p_{\text{о.д}} \geqslant p$, то кодирование явно нецелесообразно[1].

Ошибочность такого вывода проще всего показать на численном примере. Предположим, что требуется передать некоторый объем информации в

[1] В этой главе сложность кодирования и декодирования не учитывается.

1000 бит. Пусть ошибки в канале происходят независимо друг от друга с вероятностью $p=10^{-3}$. Вероятность того, что, не прибегая к помехоустойчивому кодированию, мы сможем передать верно (т. е. без единой ошибки) весь объем информации,

$$Q=(1-p)^{1000}=(1-0{,}001)^{1000}\approx 0{,}37. \qquad (13.2)$$

Предположим, что предложено использовать блочный код, в каждом блоке которого содержится 100 информационных двоичных символов, а вероятность ошибочного декодирования в данном канале $p_{\text{о.д}}=10^{-2}$. Стоит ли принять это предложение?

Конечно, нет, скажет упомянутый выше «математик». Ведь $p_{\text{о.д}}>p$, так что мы только проиграем при использовании этого кода. Но не будем торопиться с выводом. Подсчитаем, какова будет вероятность Q безошибочного приема всего объема информации в 1000 бит, который теперь состоит из 10 блоков по 100 бит в каждом. Так как вероятность правильного декодирования блока равна $1-p_{\text{о.д}}$, а ошибки возникают независимо, то вероятность того, что все 10 блоков будут декодированы верно, составит

$$Q=(1-p_{\text{о.д}})^{10}=0{,}99^{10}\approx 0{,}9, \qquad (13.3)$$

т. е. значительно выше, чем в случае передачи без помехоустойчивого кодирования, так что применение предложенного кода в данном случае повышает верность приема.

13.2. Эквивалентная вероятность ошибки

Инженеры давно убедились в недопустимости непосредственного сравнения величин p и $p_{\text{о.д}}$. Уже в первых работах по использованию помехоустой-

чивого кодирования многие авторы оценивали верность приема с помощью остаточной вероятности ошибки $p_{ост}$, т. е. вероятности того, что двоичный символ после декодирования окажется принятым неверно. Попробуем применить этот подход к рассмотренному выше примеру.

При передаче кодового блока, содержащего 100 бит информации, вероятность правильно принять все символы в нашем примере равна 0,99. С вероятностью 0,01 декодированный блок содержит несколько ошибочных символов. Сколько именно?

На этот вопрос нельзя ответить, не зная в точности структуры кода. Однако если код достаточно «мощный» (а в данном случае это, вероятно, имеет место, так как блок содержит 100 информационных символов), то при появлении в канале неисправляемого сочетания ошибок еще очень много «дополнительных ошибок» вносится при декодировании. Поэтому в первом приближении можно считать, что в ошибочно декодированном блоке от четверти до половины всех информационных символов регистрируются ошибочно. При этом предположении $p_{ост} \approx (1/4 \div 1/2) p_{о.д}$ или в нашем примере $p_{ост} \approx (2{,}5 \div 5) \cdot 10^{-3} > p$.

Получилось не то, что мы ожидали. Остаточная вероятность ошибки в двоичном символе при кодировании оказалась больше исходной вероятности ошибок в канале. Это значит, что в декодированной последовательности символов будет больше ошибочных, чем в случае передачи их без кодирования. Что ж, отсюда следует, что предложенный код в данном канале использовать не следует? А как же с предыдущим расчетом, показавшим, что вероятность правильно принять 1000 бит при кодировании выше, чем без кодирования? Неужели здесь кроется опять какой-то подвох?

Нет, здесь все верно. Несовпадение оценки объясняется различными критериями. В первом случае мы потребовали, чтобы массив информации длиной в 1000 бит с большей вероятностью был принят без ошибок, и при этом кодирование оказалось полезным. Во втором же случае мы хотели просто получить как можно меньшую вероятность ошибочного приема каждого символа, и при этом подходе кодирование оказалось вредным. Отметим также, что ошибочно принятые символы в отсутствие кодирования возникают, по условию задачи, независимо друг от друга, т. е. более или менее равномерно располагаются во всем массиве информации (поток Бернулли). При кодировании же декодированные ошибочные символы группируются в пределах отдельных ошибочно декодированных блоков.

Именно поэтому, хотя при кодировании ошибок оказалось больше, промежутки между ошибками (т. е. интервалы безошибочного приема) в среднем длиннее, чем без кодирования.

Так, какой же критерий вернее? На этот вопрос однозначно ответить нельзя. Все зависит от назначения системы связи и от свойств источника сообщения. Если источник выдает сообщения с большой избыточностью, то обычно можно считать допустимым некоторое количество ошибочно принятых символом. В первом приближении верность приема можно оценить остаточной вероятностью ошибок — чем она меньше, тем точнее восстанавливается переданное сообщение. Примером может служить синтетическая (вокодерная) телефония, телефония при импульсно-кодовой или дельта-модуляции и т. п. Здесь за счет большой избыточности речи сообщение остается разборчивым при некотором числе ошибочно принятых символов. В этом слу-

чае применять код, увеличивающий $p_{ост}$, конечно, не следует.

Иначе обстоит дело, когда передаваемые сообщения не имеют избыточности или когда эту избыточность не удается использовать для восстановления сообщения. Тогда всякая ошибка искажает принятое сообщение и во многих случаях обесценивает его полностью. Ясно, что в этих условиях критерием верности является вероятность безошибочного приема всего сообщения. При таком подходе для нас безразлично, будет ли в принятом сообщении одна ошибка или их будут сотни. Все равно сообщение искажено и не может быть использовано. В нашем примере, если длина сообщения равна 1000 бит, то из (13.2) и (13.3) видно, что применение кодирования целесообразно.

Но здесь возникает вопрос, что считать длиной сообщения, если система связи работает непрерывно. Это опять-таки определяется свойствами сообщения. Обычно всегда его можно разделить на некоторые отрезки, имеющие самостоятельное значение, так что ошибочный прием одного из них не влияет на использование остальных. Так, при телеграфной связи такими отрезками являются отдельные телеграммы. При передаче данных обычно также можно выделить отдельные массивы информации, независимые по своему целевому назначению.

Для того чтобы по возможности полнее характеризовать верность некоторой дискретной системы связи одним числом, необходимо исключить зависимость от длины сообщения. Таким числом, удобным для систем, в которых требуется безошибочный прием сообщений, является эквивалентная вероятность ошибки $p_э$. Напомним, что под этим понимается вероятность ошибки в двоичном симметричном канале с независимыми ошибками

(ДСК), обеспечивающем без применения избыточного кодирования ту же асимптотическую вероятность безошибочного приема сообщения. Другими словами, если в системе связи $Q(k)$ — вероятность безошибочного приема сообщения, содержащего k бит информации, то $p_э$ определяется из выражения

$$\lim_{k\to\infty}(Q(k)/(1-p_э)^k = 1. \qquad (13.4)$$

Приведем несколько простых примеров. Для системы связи в двоичном симметричном канале с вероятностью ошибки p, в которой используется блочный (n, k)-код, позволяющий исправлять t ошибок, эквивалентная вероятность ошибки вычисляется без предельного перехода. Как легко убедиться, она равна

$$p_э = 1 - \left[\sum_{i=0}^{t} C_n^i\, p^i (1-p)^{n-i}\right]^{1/k}. \qquad (13.5)$$

Для системы связи, не использующей помехоустойчивого кодирования, в симметричном m-ичном канале без памяти

$$p_э = p/\log m, \qquad (13.6)$$

где p — вероятность ошибки в m-ичном канале.

Вообще для систем, работающих в канале без памяти, эквивалентную вероятность ошибки можно вычислить без предельного перехода.

Обычно утверждают, что эквивалентная вероятность ошибки хорошо характеризует только те дискретные системы связи, в которых требуется безошибочный прием длинных сообщений. Однако можно обобщить это понятие, введя эквивалентную вероятность ошибки при заданном критерии верности. Пусть, например, принятое сообщение (последовательность двоичных символов) может считаться приемлемым, если доля ошибочно принятых символов не превышает ε. Тогда эквивалентная ве-

роятность ошибки $p_э(\varepsilon)$ при этом критерии верности определяется той же ф-лой (15.4) с заменой $Q(k)$ на $Q_\varepsilon(k)$ — вероятность того, что среди k принятых символов содержится не более εk ошибочных. Заметим, что в эквивалентном ДСК требуется *безошибочный* прием сообщений той же длины. При всяком другом определении не удается избежать противоречий. С увеличением допуска ε эквивалентная вероятность ошибки, при прочих равных условиях, уменьшается, что вполне естественно.

Можно распространить это понятие даже на системы передачи непрерывных сообщений. Пусть $H_\varepsilon(T)$ — эпсилон-энтропия (в битах) сообщения длительностью T, т. е. минимальное количество информации, которое необходимо передать по каналу связи для того, чтобы восстановить сообщения с данным критерием верности ε. Здесь ε может определяться как угодно, например через среднюю квадратичную погрешность, через артикуляцию и т. д. Если $Q_\varepsilon(T)$ — вероятность того, что в данной системе связи принятое сообщение удовлетворяет установленному критерию, то ф-ла (13.4) естественно преобразуется в следующую:

$$\lim_{T\to\infty}\left(Q_\varepsilon(T)/(1-p_э)^{H_\varepsilon(T)}\right)=1. \qquad (13.7)$$

Таким образом, эквивалентная вероятность ошибок является весьма универсальным параметром для сравнения и сопоставления систем связи по их помехоустойчивости. Странно, что очень долго это понятие не находило широкого применения, да и сейчас многие инженеры его не используют. Автору пришлось, например, в течение пяти лет убеждать в разумности этого понятия одного очень та-

лантливого исследователя, который предпочитал пользоваться остаточной вероятностью ошибки, гораздо хуже характеризующей верность приема.

13.3. Где применять разнесенный прием?

Однажды рецензенту одного журнала прислали статью о целесообразности применения m-ичных кодов в системах передачи дискретных сообщений. Тот факт, что переход от двоичного кода к m-ичному при $m>2$, вообще говоря, увеличивает верность, общеизвестен и, в частности, был показан в [23]. Но автор статьи ставил вопрос, в каких случаях этот переход дает больший эффект — в канале без замираний или в канале с замираниями. Статья была в целом написана грамотно. Сравнение систем с двоичным и m-ичным кодом производилось при одинаковой скорости передачи информации, одинаковых мощностях сигнала и одинаковой спектральной плотности белого шума, по эквивалентной вероятности ошибок.

Как показал расчет, в канале без замираний при некогерентном приеме ортогональных сигналов переход системы с $m=2$ к $m=32$ позволяет, например, снизить эквивалентную вероятность ошибки с $8\cdot10^{-2}$ до $1{,}5\cdot10^{-4}$, т. е. почти в 600 раз. В канале же с рэлеевскими замираниями такое мероприятие снизит эквивалентную вероятность ошибки с $8\cdot10^{-2}$ примерно до $1\cdot10^{-2}$, т. е. всего лишь в 8 раз. На этом основании в статье рекомендовалось применять $m>2$ в первую очередь для каналов без замираний, где такое мероприятие дает наибольший эффект.

Статья была забракована на том основании, что эта рекомендация неверна, по крайней мере для радиосвязи, о которой и шла речь. Прежде чем перейти к объяснению такого заключения, по-

ставим другой вопрос — где лучше применять прием на разнесенные антенны — в канале без замираний или в канале с замираниями?

Именно этот вопрос рецензент задал автору статьи, когда он приехал к нему объясняться по поводу рецензии. Он ответил, не задумываясь, так же, как ответил бы любой опытный радиоинженер. Конечно, разнесенный прием следует применять в каналах с замираниями. В каналах без замираний его никто никогда не применяет, да и вообще основная идея разнесенного приема заключается в уменьшении вредного действия замираний.

Совершенно верно, ответил рецензент. Ну, а теперь займемся небольшим сравнением.

Сравним одиночный и сдвоенный приемы на разнесенные антенны двоичных ортогональных сигналов. Прием будем для упрощения расчетов считать некогерентным и используем метод квадратичного сложения. Ветви разнесения будем считать статистически однородными, а аддитивную помеху — белым гауссовским шумом.

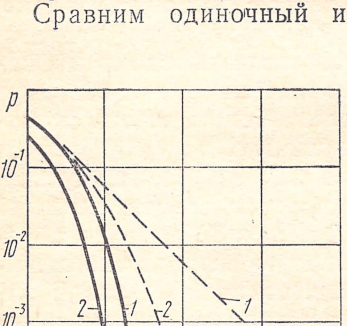


Рис. 13.1. Вероятности ошибок при одиночном *(1)* и сдвоенном *(2)* некогерентном приеме в канале:
———— без замираний;
— — — — с рэлеевскими замираниями

На рис. 13.1 представлена зависимость вероятности ошибки от среднего отношения h_0^2 энергии сигнала к спектральной плотности шума при одиночном и сдвоенном прие-

ме для канала без замираний и с рэлеевскими замираниями. Эти кривые заимствованы из [18]. Из них видно, что применение разнесенного сдвоенного приема в канале без замираний позволяет, например, снизить вероятность ошибки с $1{,}5 \cdot 10^{-2}$ до $1 \cdot 10^{-3}$, т. е. в 15 раз. В канале с рэлеевскими замираниями те же меры уменьшают вероятность ошибки с $1 \cdot 10^{-2}$ до $1 \cdot 10^{-3}$, т. е. только в 10 раз.

Что же получилось? Разнесенный прием еще более эффективен в канале без замираний, нежели в канале с рэлеевскими замираниями. Куда же смотрят инженеры? Почему они упорно применяют разнесенный прием в каналах с замираниями и игнорируют возможность получить 15-кратный выигрыш, используя разнесенный прием в канале без замираний, например, для связи на метровых волнах в пределах прямой видимости?

Успокойтесь, читатель, инженеры поступают правильно. Рассмотрим этот вопрос с позиций разработчика дискретной системы радиосвязи.

Пусть необходимо обеспечить вероятность ошибки не выше 10^{-4} в канале без замираний. Взглянув на те же кривые, легко усмотреть, что применение сдвоенного приема позволит сэкономить 3 дБ на величине h^2_0. Это значит, что можно будет применить либо передатчик с вдвое меньшей мощностью, чем при одиночном приеме, либо более дешевый приемник с коэффициентом шума, увеличенным на 3 дБ, либо немного более простую передающую антенну с вдвое меньшим коэффициентом усиления[1]. В подавляющем большинстве случаев ни одно из этих упрощений не окупит расходы по установке двух разнесенных приемных антенн,

[1] Здесь предполагается, что помехи в основном определяются шумами приемника.

двух приемников и устройства сложения. Поэтому всякий разумный разработчик откажется от варианта с разнесенным приемом.

Пусть теперь ту же верность требуется обеспечить в канале с рэлеевскими замираниями. Теперь энергетический выигрыш сдвоенного приема, как видно из рис. 13.1, составляет 17 дБ. Это позволит, например, применить в 50 раз менее мощный передатчик, скажем в 200 Вт вместо 10 кВт. С такой экономией уже нельзя не считаться. Она безусловно окупит расходы по устройству сдвоенного приема. В некоторых ситуациях, когда аппаратура должна быть легкой и компактной, только разнесенный прием позволяет обеспечить требуемую верность. Вот почему он широко используется в диапазоне декаметровых волн, в котором обычно встречаются глубокие замирания.

Здесь полезно заметить, что энергетический выигрыш разнесенного приема в 17 дБ вовсе не означает, что после сложения принятых сигналов отношение сигнал/помеха действительно увеличится на 17 дБ. Согласно известной теореме Бреннана [3] сдвоенный прием при оптимальном сложении позволяет увеличить это отношение только на 3 дБ. Энергетический выигрыш при разнесенном приеме показывает, какому увеличению этого отношения соответствует уменьшение вероятности ошибок, которое при замираниях определяется в основном уменьшением дисперсии уровня сигнала. Поэтому энергетический выигрыш (хотя и несколько меньший) имеет место и при разнесенном приеме по схеме автовыбора, когда никакого реального увеличения отношения сигнал/помеха не происходит. Этот вопрос хорошо освещен в [38].

Вернемся теперь к вопросу о целесообразности применения m-ичных кодов. Расчет показывает,

что при эквивалентной вероятности ошибок порядка 10^{-4} переход от $m=2$ к $m=32$ в канале без замираний дает энергетический выигрыш примерно 5 дБ, тогда как в канале с рэлеевскими замираниями он достигает почти 10 дБ. Здесь разница, конечно, не столь велика, как в случае разнесенного приема, но все же эффективность повышения основания кода в канале с замираниями выше, чем в канале без замираний, что диаметрально противоположно выводу, сделанному в неудачной статье, о которой говорилось в начале этого раздела.

13.4. Критерии верности телефонного сигнала

В большинстве теоретических работ о передаче непрерывных сообщений используют оценку верности приема по средней квадратичной ошибке. Иногда оговаривают, что критерий средней квадратичной ошибки используется за неимением лучшего. Его обосновывают часто тем, что он пригоден для любых видов непрерывных сообщений, а также тем, что телефонный сигнал, принятый с малой средней квадратичной ошибкой, будет наверняка разборчив. Все же основным достоинством этого критерия является лишь то, что он сравнительно просто поддается расчету.

Тем не менее средняя квадратичная ошибка плохо характеризует качество принятого слухового сигнала. Стремление ее минимизировать приводит к тому, что рассчитанные оптимальные функциональные схемы для приема модулированных звуковыми сообщениями сигналов часто содержат такие элементы, как регулируемые усилители, компенсаторы задержки, устройства подстройки фазы и т. п., которые в реальной аппаратуре обычно не применяют.

Пусть передается непрерывное сообщение $x(t)$, которое будем полагать нормированным так, что

$$\langle x^2(t)\rangle = 1, \qquad (13.8)$$

где угловые скобки означают усреднение по ансамблю. Если обозначить принятое сообщение $y(t)$, то средний квадрат ошибки ε^2, по которому оценивается верность,

$$\varepsilon^2 = \langle [x(t) - y(t)]^2 \rangle. \qquad (13.9)$$

Обычно требуют, чтобы ε^2 не превышало нескольких сотых.

Покажем, что критерий минимума ε не адекватен условиям телефонной связи. Пусть, например, из-за избыточного усиления в тракте $y(t) = 2x(t)$, т. е. принимается неискаженный по форме телефонный сигнал, но с уровнем на 6 дБ выше переданного. Естественно полагать такое изменение несущественным. Однако при этом

$$\varepsilon^2 = \langle [x(t) - 2x(t)]^2 \rangle = \langle x^2(t) \rangle = 1, \qquad (13.10)$$

что означает очень сильное искажение.

Аналогичный результат получится, если $y(t) = x(t-\tau)$, т. е. принятое сообщение запаздывает относительно переданного на некоторую величину τ, превышающую интервал корреляции $x(t)$. Ясно, что запаздывание даже на несколько десятков миллисекунд никак не отражается на качестве телефонной связи, однако по ф-ле (13.9) величина ε может оказаться даже больше единицы.

Некоторые специалисты, которым автор говорил об этом несоответствии критерия (13.9), отвечали, что в этой формуле следует под $y(t)$ понимать нормированный принятый сигнал, да еще при скомпенсированном запаздывании. Однако об этом, кажется, нигде не написано.

Но этого еще мало. Как уже отмечалось в гл. 9, сдвиг начальных фаз всех составляющих сигнала на некоторый угол φ не влияет на слуховое восприятие. Это уже никак не подразумевается в ф-ле (13.9).

Можно сформулировать критерий, значительно более адекватный телефонной связи, а именно

$$\varepsilon^2 = \min_{\varphi, \tau, a} <[ax(t-\tau)\cos\varphi + \hat{ax}(t-\tau)\sin\varphi - y(t)^2]>, \tag{13.11}$$

где $\hat{x}(t)$ — преобразование Гильберта от $x(t)$. Это значит, что перед вычитанием принятого сообщения из переданного его фаза поворачивается на φ, он сдвигается по времени на τ и умножается на множитель a, причем величины φ, τ и a выбираются так, чтобы полученное в результате сообщение как можно меньше (в среднеквадратичном смысле) отличалось от принятого. Можно при этом задать допустимые границы изменения параметров τ и a. Оставшаяся после этих операций средняя квадратичная погрешность значительно лучше характеризует качество приема телефонного сигнала.

Заметим, что изменения a, τ и φ, допустимость которых легла в основу ф-лы (13.11), не только не влияют на разборчивость, но вообще не ощущаются слухом как искажение. Необходимость ограничения значений a вызвана тем, что очень тихий, равно как и очень громкий, сигнал для получателя нежелателен. Ограничение допустимой задержки τ требуется лишь при двусторонней телефонной связи, так как при задержках свыше десятых долей секунды нарушается ощущение непосредственного контакта с собеседником. Величина φ может принимать любые значения от 0 до 2π.

Возможны и другие искажения формы сигнала, практически не влияющие на разборчивость, например не очень сильные линейные искажения, нарушающие распределение мощности в спектре сигнала. Однако они воспринимаются слухом как искажения, хотя и не затрудняющие понимания речи. Поэтому они учитываются ф-лой (13.11) только своим влиянием на ε так же, как и другие, например нелинейные, искажения и помехи.

Автор вовсе не претендует на то, чтобы ф-ла (13.11) была принята за основу при оценке верности в телефонии. Вероятно, отыскивать по ней величину ε экспериментально будет нелегко. Но возможно, что в теоретических работах она смогла бы найти применение. Ее ценность была бы доказана, если бы на ее основе удалось синтезировать более простые оптимальные схемы, чем на основе (13.9).

При экспериментальной оценке телефонных систем связи критерием средней квадратичной ошибки обычно не пользуются. Вместо этого оценивают линейные искажения в системе путем снятия частотных характеристик и определения уровня шума, а нелинейные — путем измерения коэффициентов гармоник, а иногда и комбинационных частот. Допустимые нормы при этом устанавливают на основе опыта и закрепляют в официальных документах.

Более полная экспериментальная оценка систем, предназначенных для передачи речи, осуществляется путем артикуляционных испытаний. В принципе они вполне соответствуют измерениям вероятности ошибки в дискретных системах связи. Основное отличие заключается в том, что в роли решающей схемы при артикуляционных испытаниях выступает человек, поскольку моделировать

процесс распознавания речи пока не удается. Для того чтобы исключить индивидуальные особенности, используют бригады артикулянтов.

Осмысленная речь содержит очень большую избыточность, позволяющую правильно воспринять передаваемый текст даже при очень плохом воспроизведении сигнала. Поэтому вероятности правильного приема осмысленных фраз и даже отдельных слов в более или менее удовлетворительных системах связи близки к единице и для их точной оценки пришлось бы передавать огромные массивы текста. Значительно точнее и быстрее проводятся артикуляционные испытания, если устранить или существенно уменьшить избыточность текста. Для этого чаще всего передаются таблицы отдельных слогов, не имеющих смысла, составленные так, чтобы по возможности их сочетания не вызывали никаких ассоциаций.

Вероятность правильного приема слога, выраженная в процентах, называется слоговой артикуляцией или коэффициентом разборчивости. Она играет роль, подобную вероятности правильного приема символа в дискретных системах. Аналогично определяются артикуляции слов и артикуляция фраз, соответствующие вероятности правильного декодирования блока в дискретных системах. Вследствие большой избыточности речи артикуляции слов, а тем более фраз значительно выше артикуляции слогов, так же как вероятность правильного декодирования блока при избыточном кодировании часто превышает вероятность правильного приема символа.

Например, при слоговой артикуляции 80% артикуляция слов достигает 99%, а артикуляция фраз (смысловая артикуляция) очень близка к 100%[1].

Конечно, смысловая артикуляция очень зависит от ансамбля фраз, используемых при испытаниях. Если этого не учитывать, то можно сильно ошибиться в оценке качества системы. Очень показателен в этом отношении случай, рассказанный автору Б. П. Асеевым и пересказанный в следующем разделе.

---

[1] Для расчетов электроакустических систем часто используют артикуляцию формант — величину, которая при артикуляционных испытаниях непосредственно не измеряется, а определяется пересчетом.

13.5. Чем кормить кошку?

Борис Павлович Асеев (1901—1956) был выдающимся ученым, одним из пионеров советской радиотехники. По его книгам училось не одно поколение инженеров связи. Однажды, в конце 30-х годов, Борис Павлович возглавлял комиссию по приемке одной важной разработки, существенной частью которой являлась радиотелефонная система. Разработанная аппаратура успешно выдержала испытания по всем показателям, кроме слоговой артикуляции, которая оказалась значительно ниже оговоренной в технических условиях. Комиссия пришла к решению забраковать на этом основании систему и вернуть ее на доработку.

Разработчики, естественно, возмутились этим решением и стали доказывать, что слоговая артикуляция — это надуманный теоретиками параметр, а важна лишь разборчивость осмысленной речи, которая в их системе удовлетворяет всем разумным требованиям. Для доказательства один из разработчиков схватил телефонную трубку, вызвал своего коллегу, находящегося на другом конце, и между ними произошел следующий разговор:

— Алло, Иван Петрович! Как вы меня слышите? Раз, два, три, четыре, пять. Как слышите, как слышите?

— Слышу вас отлично. А как вы меня слышите?

— И я вас слышу отлично. Раз, два, три, четыре, пять. Вы все понимаете?

— Да, я вас отлично понимаю.

— Ну, вот видите, Борис Павлович, разборчивость превосходная, а ваши артикуляционные испытания просто вводят в заблуждение.

— Позвольте-ка мне поговорить, — сказал Бо-

рис Павлович своим хорошо поставленным басом и взял у него трубку.

— Алло, алло, Иван Петрович, с вами говорит Асеев. Как вы меня слышите?

— Слышу вас прекрасно, Борис Павлович, разборчивость полная.

— А скажите, Иван Петрович, вы когда-нибудь вашу кошку грушами кормили?

— Что?

— Вы кошку грушами кормили?

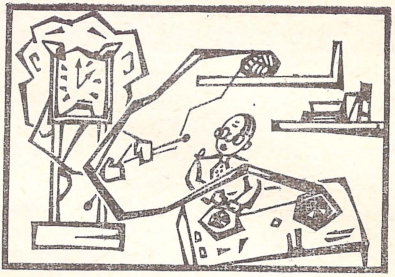
— Не понимаю. Повторите еще раз.

— Я спрашиваю, вы вашу кошку грушами кормили?

Но сколько ни повторял Борис Павлович свой вопрос, стараясь произносить слова внятно и медленно, его собеседник ничего понять не мог. После этого эксперимента разработчики уже не пытались возражать против решения комиссии.

Разумеется, набор стандартных фраз, вроде «как вы меня слышите?», представляет множество со значительно большей избыточностью, нежели обычный телефонный разговор. Поэтому даже при очень низкой слоговой артикуляции такие стандартные фразы «декодируются» безошибочно. Но стоило выйти за пределы стандартного набора и задать вопрос хотя и осмысленный, но довольно необычный, как артикуляции не хватило.

На этом воспоминании мы закончим разговор о методах сравнения систем связи, которые, пожалуй, заслуживают отдельной монографии.



14

НЕМНОГО ИНФОРМАЦИИ

14.1. О постулатах теории информации

Теория информации зародилась как математическая теория связи. Именно так называлась статья Шеннона [56], в которой впервые были сформулированы основные понятия и теоремы. Конечно, понятие «связь» можно трактовать в очень широком смысле, что и было сделано многими исследователями в различных областях науки, например лингвистике, теоретической физике, экономике и т. п. В ряде случаев дело сводится к механическому переносу терминов теории информации, иногда без ясного понимания их смысла. Возникла опасность дискредитации основных идей теории информации.

Первым, забившим тревогу по этому поводу, был сам Шеннон [59]. После его выступления поток публикаций, в которых понятия теории информации трактуются весьма расширительно, несколько поредел. Однако и до сих пор время от времени появляются работы, в которых идеи теории информации применяются к самым различным проблемам точных, естественных и гуманитарных наук. Следует признать, что многие из этих работ содер-

жат весьма полезные и достаточно обоснованные результаты.

Наряду с этим появилось немало работ, в которых подвергаются критике некоторые уже ставшие классическими приложения теории информации, в том числе и к вопросам теории связи.

Оставив в стороне содержание этих работ, попытаемся четко сформулировать условия (постулаты), при которых в теории информации исследуются системы связи. Некоторые из этих условий обычно подразумеваются, но далеко не всегда указываются в явном виде. Для облегчения задачи ограничимся только передачей дискретных сообщений:

1. Источник сообщения осуществляет выбор сообщения из некоторого множества с определенными вероятностями.

2. Сообщения могут передаваться по каналу связи в закодированном виде. Кодированные сообщения образуют множество, являющееся взаимно однозначным отображением множества сообщений[1]. Правило декодирования известно декодеру (записано в его программе).

3. Сообщения следуют друг за другом, причем число сообщений, от которых зависит кодовый символ (длина кодовых ограничений), может быть сколь угодно большим.

4. Сообщение считается принятым верно, если в результате декодирования оно может быть в точности восстановлено. При этом безразлично, сколько времени прошло с момента передачи сообщения до момента окончания декодирования, а также ка-

---

[1] Вообще говоря, кодирование может производиться и неоднозначно. Важно, чтобы однозначно происходило декодирование (в отсутствие шума).

кова сложность операций кодирования и декодирования.

5. Количество информации не зависит от смыслового содержания сообщения, от его эмоционального воздействия, от его полезности и даже от его отношения к реальной действительности.

Из этих условий только первое сформулировано четко в работах Шеннона, а также в большинстве монографий о теории информации. Однако легко убедиться, что нарушение любого из них либо не позволяет определить количество информации, либо существенно влияет на смысл теорем кодирования Шеннона. Рассмотрим их по порядку.

14.2. Первые три условия

На возможность нарушения первого условия и на затруднения в определении при этом количества информации, по-видимому, впервые указал академик А. Н. Колмогоров [20]. В частности, он писал: «...Практически можно считать, например, вопрос об «энтропии» потока поздравительных телеграмм... корректно поставленным в его вероятностной трактовке и при обычной замене вероятностей эмпирическими частотами... Но какой реальный смысл имеет, например, говорить о «количестве информации», содержащейся в тексте «Войны и мира». Можно ли включить разумным образом этот роман в совокупность «возможных романов», да еще постулировать в этой совокупности некоторого распределения вероятностей?...»

Существуют и другие ситуации, в которых трудно говорить о каких-либо вероятностях. В работе [61], содержащей немало глубоких мыслей, отмечается, что нельзя задать вероятность появления

самолета над данным районом. Поэтому нельзя определить и энтропию сообщения, выдаваемого радиолокационной станцией обнаружения. Вряд ли с этим утверждением можно согласиться (см. ниже).

В упомянутой работе А. Н. Колмогорова намечаются некоторые пути определения количества информации без использования понятия вероятности. К сожалению, эти идеи не доведены до конца.

Второе условие обычно выполняется в системах связи. Однако существуют ситуации, когда правило декодирования получателю неизвестно. Одна из них возникает при приеме шифрованного сообщения, если принцип шифрования или хотя бы использованный ключ неизвестны. Этому вопросу, между прочим, посвящена работа Шеннона [57]. Другой пример относится к связи с инопланетными цивилизациями. Сможем ли мы обмениваться с ними информацией, не договорившись заранее о коде?

В 1967 г. в обсерватории Кембриджского университета при изучении космических радиоизлучений на частотах порядка 100 МГц обнаружили периодическую последовательность коротких радиоимпульсов, приходящих из одной точки Галактики. Эти импульсы имели чрезвычайно стабильный период следования, немного больше 1 с. Они настолько отличались от обычного излучения космических объектов, имеющих характер гауссовского шума с неравномерным спектром, и так походили на сигналы наших земных радиомаяков, что в первый момент обнаруживший их проф. Хьюиш не сомневался в том, что это сигналы, посылаемые разумными существами. Именно поэтому он не торопился с публикацией и продолжал изучать эти импульсы. Оказалось, что таких источников радио-

импульсов (пульсаров) существует довольно много и вскоре появилось несколько гипотез об их природе, объяснявших периодичность и другие особенности импульсов без предположения об участии в этом инопланетных цивилизаций. В настоящее время считается, что пульсары — это быстро вращающиеся нейтронные звезды.

Но давайте пофантазируем. Предположим, что обнаружено космическое излучение, представляющее последовательность импульсов (единиц) и пауз (нулей), например, такого характера:

0001001000110100010101100111110001001 и т. д.

На первый взгляд это довольно хоатическая последовательность. Однако, разбив эти символы на группы по четыре, легко обнаружить, что они представляют двоичную четырехразрядную запись последовательности натуральных чисел: 1, 2, 3, 4, 5, 6, 7, 8, 9 и т. д. Можно ли утверждать, что это излучение посылается разумными существами?

Оценить вероятность такой гипотезы практически невозможно, хотя бы потому, что об априорных вероятностях существования цивилизации в некотором участке Галактики нам в сущности ничего неизвестно. Можно, однако, говорить о правдоподобии этой гипотезы, которое довольно велико, так как если предположить существование инопланетных разумных существ, которые хотят установить контакт с «собратьями по разуму», то с большей вероятностью они пошлют именно такого рода сигналы.

Вполне возможно, что будет высказана и противоположная гипотеза о «неразумном» происхождении этого излучения, но если не будет найдено никакого правдоподобного объяснения возникновения такой последовательности импульсов в результате

физических процессов, то можно будет утверждать, что мы извлекли из кодированного сигнала информацию о существовании внеземной цивилизации, не договорившись заранее о коде.

В [34] рассказывается об эксперименте, проведенном на радиоастрономической конференции в Грин Бэнк (США), где участникам роздали ленты с последовательностью нулей и единиц и предложили расшифровать эту «запись сигнала из космоса». Большинство участников успешно справились с этой задачей, догадавшись, что они имеют дело с разверткой изображения на плоскости. Восстановив это изображение и проявив еще много догадливости, они могли судить об облике существ, пославших сигнал, об их планетной системе и о многом другом. Это тоже подтверждает, что, не договариваясь о коде, можно извлекать информацию из сигнала.

Еще примеры — исследуя спектр излучения звезды, мы извлекаем информацию о ее химическом составе, температуре, скорости движения относительно солнечной системы, магнитном поле и т. д. И все это тоже без договоренности о коде.

Таким образом, для получения информации в принципе можно и не знать, как закодировано сообщение, если этот код можно раскрыть. Всегда ли это возможно? На это автор ответить не берется. Лично он думает, что не всегда. Факт «раскрытия» кода спектральных линий не может служить опровержением, так как они наблюдались и в лабораторных условиях, когда источник излучения и его физическое состояние были хорошо известны. Раскрытие генетического кода, которое считают одним из крупнейших достижений современной науки, облегчалось тем, что наряду со структурой молекулы дезоксирибонуклеиновой кислоты (ДНК)

исследователи наблюдали и синтезированный с ее помощью белок.

Итак, можно считать, что для применимости теорем кодирования теории информации в общем случае требуется наличие у декодера предварительной информации о коде.

Третье условие также совершенно необходимо для применимости теорем кодирования, поскольку они имеют асимптотический смысл. К автору неоднократно обращались за советами такого характера: как закодировать сообщение, содержащее всего лишь 10 бит, чтобы оно было принято верно с вероятностью $1 \cdot 10^{-6}$, если вероятность ошибки в канале p равна 10^{-2}, а передать можно не более 20 двоичных символов. В ответ автор только разводил руками и признавался в своем бессилии.

— Как же так? — возмущался спрашивающий.— Ведь пропускная способность двоичного симметричного канала равна $1+p\log p+(1-p)\log(1-p)$, что составляет при $p=10^{-2}$ более 0,9 бит на символ. Следовательно, по теореме кодирования достаточно иметь хоть немного более 10% избыточных символов, чтобы декодировать со сколь угодно малой вероятностью ошибки. А я допускаю даже 50% — 10 избыточных символов из 20-ти. Вероятно, вы просто не знаете хороших кодов.

Что можно на это возразить? Конечно, если бы длина кодового блока была 100 при 50 информационных символах, то задача бы легко решалась. В данном же случае, если даже выторговать у заказчика лишний проверочный символ, лучшее что можно применить — это укороченный код Голлея (21, 10), который в этих условиях для канала без памяти позволяет получить вероятность ошибочного декодирования около $5 \cdot 10^{-5}$, т. е. в 50 раз больше допустимой. Еще хуже будет в канале с па-

мятью, хотя его пропускная способность и выше, так как в этом случае асимптотика проявляется лишь при длине кода, превышающей интервал сильной корреляции между ошибками.

14.3. Информация и время

В четвертом условии самым важным является возможность задерживать сообщения на сколь угодно большое время. Строго говоря, это условие на практике никогда не выполняется. Даже поздравительная телеграмма, пришедшая через неделю после дня рождения, не доставит радости получателю. Правда, такие задержки вызываются обычно не процессами кодирования и декодирования, а другими причинами. Однако бывают нередко случаи, когда и малые задержки недопустимы, что накладывает ограничения на длину кода и не позволяет в полной мере применить теоремы кодирования. Это возможно, например, при передаче информации в автоматизированных системах управления различными технологическими процессами. В этих случаях обычно говорят о передаче в реальном масштабе времени. Но этот термин никак не вписывается в понятия теории информации.

Следует особо отметить один вид сообщений, к которому теория информации, по-видимому, не применима из-за нарушения четвертого условия. Это — сообщения, относящиеся к самому времени. Представьте себе, что с сигналами точного времени, передаваемыми в сети радиовещания, поступили так же, как обычно поступают с различного рода радиорепортажами, музыкальными передачами и т. п., т. е. сначала записали на пленку, а потом воспроизвели. Конечно, такие сигналы никакой информации о времени уже не содержат. Поэтому

нельзя говорить о каком-то кодировании сигналов времени. То же самое (хотя и с некоторыми оговорками, относящимися к периодичности) можно сказать о передаче информации, необходимой для синхронизации каких-либо процессов, в частности самого процесса обработки и декодирования сигналов. И хотя нетрудно вычислить количество информации, необходимое для осуществления синхронизма с заданной точностью, передавать ее необходимо с учетом ее специфики. Это, конечно, не относится к «цикловой синхронизации», задачей которой является не привязка к определенному моменту времени, а указание на первый символ в кодовом блоке.

Интересно с этой точки зрения рассмотреть условия передачи информации об обнаружении самолета радиолокационной станцией. Автор работы [61] полагает, что невозможность задать вероятность появления самолета над данным районом в данный отрезок времени препятствует применению теории информации в данном случае. Однако всегда можно приблизительно оценить эту вероятность снизу и определить, насколько можно сжать передаваемую информацию. Пусть, например, в условиях военных учений самолет появляется в среднем 1 раз в 3 ч, а фиксировать момент появления нужно с точностью до 1 с. Тогда вероятность появления самолета в данном секундном интервале $p<10^{-4}$ и энтропия источника не превышает величины $H=-P\log P-(1-P)\log(1-P)<1{,}2\times\times10^{-3}$ бит. Такое количество информации требуется передавать в каждую секунду.

В сутках содержится 86 400 с. Пронумеруем их и будем в конце каждых суток передавать номера секунд, в которые отмечалось появление самолета. Для передачи каждого номера потребуется $\log 86\,400<17$ бит. В среднем за сутки самолеты появляются 8 раз или с учетом возможного разброса до 12 раз. Следовательно, можно передать информацию обо всех появлениях самолетов за истекшие сутки кодовой последовательностью, содержащей 204 двоичных символа. В пересчете на секунду это получится около $2{,}3\cdot10^{-3}$ символа, что превышает энтропию меньше чем вдвое. Можно еще лучше сжать информацию и приблизиться к предельному значению скорости передачи, если посылать сообщения сразу за десять

суток. Беда лишь в том, что эта задержанная информация никакой ценности не представляет — о появлении самолета необходимо сообщить немедленно. Итак, невозможность сжать информацию, передаваемую от радиолокационной станции, обусловливается нарушением не первого, а четвертого условия.

Интересно было бы решить такую задачу. По каналу с шумом передаются измеренные значения некоторой функции времени. Получатель должен знать значения, которые принимает функция, в каждый данный момент без задержки. Если передавать сигналы без помехоустойчивого кодирования, то задержка будет очень мала и, экстраполируя принятый сигнал, получатель смог бы хорошо оценивать нужные ему значения функции, если бы не помехи, вносящие заметные ошибки. Для защиты от этих ошибок можно ввести помехоустойчивое кодирование. Но тогда задержка сигнала значительно увеличится, экстраполировать придется на больший интервал времени и появятся ошибки экстраполяции. Требуется найти оптимальный метод кодирования, минимизирующий суммарную ошибку от помех и экстраполяции. Решение, конечно, должно зависеть от характеристик канала и от корреляционной функции передаваемого сообщения. Оно представило бы серьезный вклад в построение раздела теории связи, посвященного передаче стареющей информации.

14.4. Семантическая информация

Пятое условие относится не к применимости теорем кодирования, а к самому определению количества информации. Оно является совершенно естественным для систем связи. В самом деле, инженеру, проектирующему, строящему или эксплуати-

рующему, например, телеграфную систему, необходимо обеспечить передачу всех возможных телеграмм независимо от того, радостные они или печальные, относятся ли к важнейшим политическим событиям, крупным научным открытиям или сообщают о приезде тети Маши. В задачу инженера-связиста вовсе не входит проверять, соответствует ли действительности телеграмма о том, что тов. Иванов заболел и не может вернуться в срок из отпуска. Его интересует чисто количественный аспект информации, определяющий в конечном счете минимальные энергетические затраты, время и полосу частот, необходимые для ее передачи.

Когда же понятие «информация» пытаются применить к области гуманитарных наук, в частности к языкознанию, то неизбежно всплывают вопросы о семантическом (смысловом) содержании. К счастью, в классической теории информации имеются понятия, довольно хорошо приспособленные для установления количественных соотношений с учетом смыслового содержания сообщений. К таким понятиям в первую очередь относятся взаимная информация $I(x, y)$ и условная взаимная информация $I(x, y|z)$.

Пусть Y — множество событий y, связанных между собой некоторым смысловым содержанием, а x — сообщение. Тогда $I(x, y)$ представляет собой количество информации, содержащееся в x относительно события y. Если известны вероятности $p(x)$, $p(y)$, $p(x, y)$, то $I(x, y) = \log\{p(x, y)/[p(x)p(y)]\}$. Математическое ожидание этой величины по совместному ансамблю XY называется средней взаимной информацией и обозначается $I(X, Y)$. В отличие от энтропии $H(X)$, которую можно рассматривать как среднюю информацию $I(X, X)$, содержащуюся в сообщении x от-

носительно самого себя, и которая никак не связана с смысловым содержанием x, а только с распределениями вероятностей на множестве сообщений, величина $I(x, y)$ зависит от смысла. Так, одно и то же газетное сообщение может содержать очень большую информацию о ходе чемпионата страны по футболу, очень малую информацию о погоде и никакой информации о политических событиях в мире.

Величина $I(x, y|z) = \log(p(x, y|z)/p(x|z)p(y|z))$ может иметь различную содержательную интерпретацию, в частности, ее можно трактовать как информацию, содержащуюся в сообщении x относительно событий y, при условии, что получателю доступна информация, содержащаяся в z. Математическое ожидание этой величины по совместному ансамблю XYZ называется средней взаимной условной информацией $I(X, Y|Z)$.

Важно отметить, что значение $I(X, Y|Z)$ может быть как больше, так и меньше $I(X, Y)$. Это легко показать на простых примерах. Пусть, например, $x=y=z$. Тогда, очевидно, $I(X, Y)=I(X, X)=H(X)$, тогда как $I(X, Y|Z)=0<I(X, Y)$. Другой пример — пусть $X=\{0,1\}$, $Y=\{0,1\}$, $Z=\{0,1\}$, причем $y=x+z \pmod 2$. Если x и z принимают значения 0 и 1 с равными вероятностями и x не зависит от z, то легко убедиться, что и y не зависит от x и с равными вероятностями принимает значения 0, 1. Тогда если z неизвестно, x и y независимы. Поэтому $I(X, Y)=0$. Если же значение z задано, то x определяется по y однозначно и $I(X, Y|Z)=H(X)=1$, т. е. $I(X, Y|Z)>I(X, Y)$.

Таким образом, некоторые предварительные знания могут в одних случаях увеличивать, а в других случаях уменьшать получаемую информацию. В этом отношении условная взаимная инфор-

мация отличается от условной энтропии $H(X|Z)$, которая никогда не может быть больше безусловной: $H(X|Z) \leqslant H(Z)$, причем равенство имеет место только при независимых X и Z.

К сожалению, понятия энтропии и взаимной информации иногда путают. На этой почве возникают досадные недоразумения. Так, например, в [61] доказывается, что понятия «классической» теории информации не применимы к «семантической» теории информации, изучающей вопросы извлечения смысла из сообщения. Автор исходит из очевидных положений:

«...Человек, изучавший некоторую отрасль науки, извлечет из специального текста по этой отрасли, вообще говоря, больше, чем до обучения... Возьмем в качестве примера сообщения текст учебника по теории вероятности и дадим двум лицам — приемникам информации: школьнику третьего класса... и, скажем, студенту-математику, который знает анализ, основы математики, но не изучал теории вероятностей. Кто из них извлечет из этого учебника больше информации? Ответ ясен — получит больше информации студент... Конечно, встречается и обратная картина, когда количество принимаемой информации уменьшается с увеличением априорного знания».

Все это совершенно верно и может быть легко описано в терминах теории информации. Пусть x — текст учебника, y — сумма всех современных знаний человечества о теории вероятностей, z_1 — все сведения, хранящиеся в памяти школьника, z_2 — студента и z_3 — крупного ученого, специалиста по теории вероятностей. Тогда, очевидно, $I(x, y|z_1) < I(x, y|z_2) > I(x, y|z_3)$, т. е. с увеличением объема знаний количество получаемой информации сначала возрастает, а затем убывает. И ни-

чего противоречащего классической теории информации здесь нет. Однако в [61] говорится следующее:

«Заметим, что эта особенность модели семантической информации существенно отличает ее от классической теории информации... В последней увеличение априорной информации всегда уменьшает количество информации, извлекаемой из данного сообщения».

Последнее утверждение, как мы видели, явно неверно. Объяснить его можно только тем, что автор [61] спутал взаимную информацию (информацию, *извлекаемую* из данного сообщения) с энтропией (мерой неопределенности, или максимальной информацией, которую *можно извлечь* из сообщения).

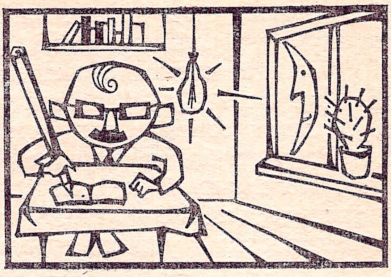
Следует отметить, что в цитированной работе [61] содержатся и другие доводы об ограниченности классической теории информации. Некоторые из них справедливы и заслуживают внимания. Однако один довод, так же как и рассмотренный выше, основан на недоразумении. Приведем еще одну цитату из той же работы:

«Существует целый ряд ситуаций, когда ясно, что для описания информации нужны характеристики ее содержания, а не статистические свойства, связанные с частотой передачи тех или иных сигналов. Это можно видеть уже из следующего факта. *В статистической теории информации чем реже источник передает данный символ, тем большее количество информации связано с этим символом* (курсив мой — Л. Ф). Однако хорошо известно, что в процессе понимания смысла текста на естественном языке редко встречающиеся элементы несут сравнительно малую долю информации. Так, хорошо известно, что содержание текста вполне

понятно человеку, владеющему лишь частью словаря данного языка, состоящей из самых частых слов, которые составляют меньшую часть языка».

Ошибочность утверждения, отмеченного курсивом, ясна всякому, знакомому с теорией информации. Действительно, пусть вероятность i-го слова в словаре равна p_i. Средняя собственная информация или энтропия на одно слово (если не учитывать связей между словами) равна $\Sigma p_i \log(1/p_i)$, где суммирование производится по всем словам языка. Вклад i-го слова в эту сумму определяется величиной $p_i \log(1/p_i)$, которая имеет максимум при $p_i = 1/e$, а с уменьшением значения p_i монотонно убывает. Так как в любом естественном языке все $p_i \ll 1/e$, то отсюда следует, что наибольшая часть информации связана с наиболее частыми словами в полном соответствии с практикой.

Мы далеки от мысли, что теория информации Шеннона может без всякого переосмысливания и дальнейшего развития применяться ко всем вопросам языкознания. Однако критика ее, содержащаяся в [61], в большей своей части не обоснована и может только ввести в заблуждение читателей.



15

МЕЛОЧИ

15.1. ФМ = ДБП?

Хорошо известно (см., в частности, гл. 7), что один и тот же модулированный сигнал $s(t) = A\cos[\omega_0 t + m\Psi(t)]$ можно рассматривать как модулированный по фазе или по частоте. Но в первом случае модулирующим (первичным) сигналом является $\Psi(t)$, а во втором случае — $x(t) = d\Psi/dt$. Рассмотрим другой вопрос — существует ли такой первичный сигнал $x(t)$, который дает одинаковый результат, если им модулировать синусоиду по фазе или по амплитуде (с подавленной несущей). Другими словами, при каком $x(t)$ возможно равенство

$$Ax(t)\cos(\omega_0 t + \varphi) = A\cos[\omega_0 t + mx(t) + \psi]? \qquad (15.1)$$

Ответ очень прост. Это равенство справедливо, если $x(t)$ принимает только два значения: $+1$ и -1, $m = \pi/2$ и $\psi = \varphi - \pi/2$. Таким образом, двоичную фазовую манипуляцию можно с таким же основанием рассматривать как «балансную манипуляцию». Отсюда, кстати, вытекает идентичность синхронного и фазового детекторов.

Заметим, что если к сигналу (15.1) добавить несущую, сохранив условие, наложенное на $x(t)$, то получится сигнал с амплитудной манипуляцией $A[1+x(t)]\cos(\omega_0 t+\varphi)$. Если в амплитудноманипулированном сигнале подавить не только несущую, но и одну из боковых полос, то получится сигнал довольно сложной формы, модулированный как по амплитуде, так и по фазе. В частности, если $x(t)$ — периодическая функция, принимающая поочередно значения $+1$ и -1 на тактовых интервалах длины T, то при однополосной модуляции получится функция $s(t)$, у которой число переходов через нуль в одном тактовом интервале с $x=1$ на два больше, чем в соседнем, с $x=-1$, если передается нижняя боковая полоса, и наоборот, если передается верхняя боковая полоса. С некоторой натяжкой такой сигнал можно назвать манипулированным по частоте.

15.2. Немного о белом шуме

Понятие «белый шум» — математическое, а не физическое, так же как и дельта-функция Дирака. Белый шум, по определению, имеет бесконечную дисперсию или, другими словами, мощность[1]. Существование источника белого шума даже со сколь угодно малой, но конечной спектральной плотностью противоречит закону сохранения энергии.

[1] Некоторые исследователи вводят понятие «белый шум с конечной мощностью», т. е. процесс, спектральная плотность мощности которого $G(\omega)$ постоянна для всех частот, а мощность, т. е. $\int_{-\infty}^{\infty} G(\omega)\,d\omega$, конечна. Это можно было бы понимать в предельном смысле, когда $G(\omega)\to 0$ при всех ω. Но ясно, что такой «белый шум» с нулевой спектральной плотностью не может быть ни помехой, ни переносчиком информации.

К сожалению, об этом часто забывают. Таким процессом с бесконечной дисперсией должно быть, с точки зрения классической (доквантовой) физики, излучение абсолютно черного тела. Именно это обстоятельство побудило Макса Планка пересмотреть представления классической электродинамики и выдвинуть квантовую теорию, постулирующую дискретность излучаемой энергии. С учетом квантовых явлений спектральная плотность мощности теплового шума определяется ф-лой (9.1), т. е. зависит от частоты и на высоких частотах быстро убывает.

Белый шум обычно определяют как стационарный процесс с корреляционной функцией $B(\tau) = (N_0/2)\delta(\tau)$, где N_0 — односторонняя спектральная плотность мощности, не зависящая от частоты.

Каковы распределения вероятностей белого шума? Этот вопрос, строго говоря, не имеет смысла. Как отмечено в [26], «распределение вероятностей белого шума в обычном смысле не существует». Почему же тогда часто говорят о гауссовском белом шуме?

Этот вопрос хорошо пояснен в [10] и [39]. Белый шум $\xi(t)$ является обобщенным процессом, который, как и обобщенная функция, получает реальный смысл только в подынтегральных выражениях функционалов, причем для всякой функции $g(t)$ из L_2 интеграл $\int_{-\infty}^{\infty} g(t)\xi(t)dt$ является по определению гауссовской величиной с нулевым математическим ожиданием и с дисперсией, равной $(N_0/2)\int_{-\infty}^{\infty} g^2(t)dt$. Отсюда следует, что

$$\eta(t) = \int_{-\infty}^{\infty} \xi(\tau)\, g(t-\tau)\, d\tau \qquad (15.2)$$

является обычным стационарным гауссовским шумом с конечной дисперсией и корреляционной функцией

$$<\eta(t)\,\eta(t+\tau)> = (N_0/2) \int_{-\infty}^{\infty} g(t)\, g(t+\tau)\, dt. \qquad (15.3)$$

Будем рассматривать стационарный процесс $\zeta(t)$ с очень малым интервалом корреляции Δ как некоторое приближение к математическому понятию «белый шум». Подадим его на вход линейной цепи с импульсной характеристикой $g(t)$. Тогда процесс $\eta(t)$ на выходе фильтра выражается ф-лой (15.2) с заменой $\xi(t)$ на $\zeta(t)$. Если интервал Δ значительно меньше протяженности импульсной характеристики фильтра $g(t)$, то на основании центральной предельной теоремы можно утверждать, что процесс $\eta(t)$ будет близок к гауссовскому шуму, каковы бы ни были распределения вероятности $\zeta(t)$. Поэтому белый гауссовский шум можно наглядно представлять как предел стационарного процесса $\zeta(t)$ при $\Delta \to 0$.

В связи с этим непонятен смысл утверждений, которые еще недавно можно было встретить в книгах, о том, что белый шум может быть не только гауссовским, но и любым другим, например рэлеевским. Сам по себе белый шум как обобщенный процесс не имеет никаких распределений вероятностей, а после «прохождения через фильтр» он всегда превращается в гауссовский процесс. Говорить о негауссовском белом шуме, по убеждению автора, бессмысленно.

15.3. Почему сигналы случайны?

Сигналы, передаваемые по каналам связи, представляют собой множество функций времени (непрерывного или дискретного) с вероятностной мерой, заданной на этом множестве. Это совпадает с определением случайного процесса. В вырожденном случае, когда множество состоит из одной функции и ее вероятность поэтому равна единице, энтропия процесса равна нулю и он не может быть переносчиком информации.

Автор просит прощения у читателя за пересказ давно известной истины. Он не стал бы этим заниматься, если бы в отдельных книгах, в том числе учебниках, не встречались бы высказывания вроде такого:

«Сущность связи состоит и в том, чтобы передать получателю неизвестные ему сведения. Сигналы, несущие такие сведения, на приемном конце заранее также будут неизвестными. Сигналы и тем более помехи для получателя являются случайными (недетерминированными).

Необходимо подчеркнуть относительность понятия недетерминированности. Сигнал для отправителя на передающем конце детерминирован, так как при заданном способе передачи он определяется известным сообщением. Для получателя тот же сигнал недетерминирован, так как передаваемое сообщение на приемном конце неизвестно» [16].

Здесь случайность трактуется субъективно. По существу, в приведенном отрывке математическое понятие «случайность» подменяется бытовым, по мнению авторов — более доходчивым смыслом этого слова: для меня случайно то, чего я сейчас не знаю. Если стать на эту точку зрения, то можно объявить любую физическую константу, например

заряд электрона, случайной величиной, поскольку я ее значения не помню. Когда же я загляну в справочник, она перестанет быть для меня случайной.

Вряд ли такое субъективистское толкование случайности полезно даже для пояснения студентам. Вероятность, как и информацию, следует рассматривать объективно. Сигнал случаен не потому, что кто-то его заранее не знает, а только потому, что он выбирается с определенной вероятностной мерой из некоторого множества. Всякая же реализация сигнала является детерминированной функцией.

15.4. Оценка энтропии текста методом отгадывания

В одной из работ Шеннона [58] обсуждается задача вычисления энтропии английского текста. По определению энтропия в расчете на одну букву равна

$$H = \langle \log (1/p) \rangle, \tag{15.4}$$

где p — вероятность буквы в данном контексте. Эта вероятность является случайной величиной, зависящей как от самого значения буквы, так и от всего предыдущего текста.

Если бы мы располагали таблицами вероятностей длинных последовательностей букв в множестве текстов для данного языка, то можно было бы вычислить энтропию по формуле, вытекающей из (15.4):

$$H = \frac{1}{n} \langle \log (1/p_n) \rangle, \tag{15.5}$$

где p_n — вероятность последовательности из n букв, а n — столь большое число, что вероятност-

ные связи между буквами дальше n не распространяются.

К сожалению, такие таблицы для различных языков существуют только при $n \leqslant 3 \div 4$, тогда как вероятностные связи простираются значительно дальше. И вот идея Шеннона заключалась в том, что эти таблицы может заменить человек, хорошо владеющий данным языком, который в ряде случаев может безошибочно определить букву, следующую за данным отрывком текста. Конечно, это не значит, что он мог бы составить таблицу вероятностей различных последовательностей букв или тем более указать алгоритм, с помощью которого он угадывает следующую букву. Многое из того, что происходит в нашем сознании, трудно или даже невозможно описать алгоритмом, например мы все еще не знаем, как опознаются различные слуховые или зрительные образы. Тем не менее, такие алгоритмы, нам неизвестные, существуют и мы ими повседневно пользуемся.

Рабочая гипотеза, выдвинутая Шенноном, заключается в том, что свободно владеющий языком человек может быть «идеальным предсказателем». Это значит, что, отгадывая следующую букву, он всегда называет ту, которая является наиболее вероятной в данном контексте. Если он ее не отгадал и ему сообщили об этом, то в следующей попытке он назовет вторую по вероятности букву и т. д.

Если теперь записать последовательность чисел, определяющих число попыток при отгадывании каждой буквы, то эта последовательность будет содержать полную информацию о тексте. Действительно, другой «идеальный предсказатель» при аналогичном испытании должен называть буквы в том же порядке, а так как полученная числовая последовательность указывает, на какой букве следует

каждый раз остановиться, то весь текст будет восстановлен полностью.

Таким образом, энтропия текста равна энтропии полученной последовательности чисел, а последнюю оценить значительно проще, хотя бы потому, что в ней значительно меньше проявляются вероятностные связи между символами (числами). Интересующихся подробностями этой оценки отсылаем к упомянутой статье [58].

Не следует думать, что использование человека в этом эксперименте, означает подмену объективного понятия вероятности субъективным, аналогичным тому пониманию случайности, которое критиковалось в предыдущем параграфе. Нет, здесь имеется в виду вероятность в строго научном, объективном смысле. Но в соответствии с гипотезой Шеннона человек, владеющий языком, умеет если не определить точно вероятности различных букв, следующих за данным отрывком, то хотя бы разместить их в порядке убывающих вероятностей. Никто никогда не утверждал, что эта гипотеза абсолютно верна, но она, по-видимому, не далека от истины (опять-таки в вероятностном смысле), и поэтому предложенная процедура годится для приближенного экспериментального измерения энтропии текста.

Многие лингвисты использовали предложенный метод для определения энтропии текстов на различных языках. К сожалению, процедура отгадывания, описанная Шенноном, в применении к достаточно длинному тексту (а это необходимо для того, чтобы учесть все дальние вероятностные связи) довольно длительна и утомительна. И вот некоторые экспериментаторы решили ее рационализировать. Вместо одного отгадывающего они использовали бригаду, например 50 человек. Каж-

дый из участников называет предполагаемую следующую букву. Относительное число голосов, поданных за данную букву, принимается за ее вероятность p. Это позволяет сразу определить значение $\log 1/p$ и, усреднив его по всем буквам в тексте, найти статистическое значение энтропии по ф-ле (15.4) с заменой математического ожидания статистическим средним. Не правда ли, прекрасная рационализация?

Так может показаться только с первого взгляда. В действительности «рационализированная» процедура ничем не обоснована. Если предположить, что все участники бригады являются идеальными предсказателями, т. е. удовлетворяют гипотезе Шеннона, то все они всегда должны называть одну и ту же букву, наиболее вероятную, даже если ее вероятность значительно меньше половины. А в соответствии с выбранной процедурой ее вероятность будет принята за единицу. Если же в бригаде имеются разногласия, то это свидетельствует лишь о том, что участники бригады не идеальные предсказатели. Но можно ли по числу их голосов судить о вероятности названных ими букв? Это было бы так, если бы можно было утверждать, что такой «неидеальный» предсказатель называет ожидаемую букву случайно, причем вероятности $p_в$ выбираемых им букв совпадают с вероятностями p появления букв в тексте. Эта гипотеза представляется мало правдоподобной. Ведь каждый из участников бригады будет называть ту букву, которая представляется ему наиболее вероятной, и в тех случаях, когда вероятность какой-то буквы заметно превышает вероятности каждой из остальных, скорее всего все назовут именно ее.

Конечно, использование бригады, в принципе, может улучшить процедуру Шеннона, однако не

при такой трактовке результатов. Правильнее было бы рассматривать решение большинства как близкое к решению «идеального предсказателя». Поэтому нужно не подсчитывать голоса, поданные за другие буквы, а отбрасывать их. Если же названная большинством буква угадана неверно, то нужно сообщить об этом и предложить всей бригаде назвать другую букву так же, как это делал Шеннон с одиночным предсказателем. Это не облегчит и не ускорит процедуру, но сделает результат более надежным, если бригада в целом ближе к «идеальному предсказателю», чем любой из ее членов.

15.5. Кое-что о терминах

Говорят, что молодые ученые не обращают внимания на чистоту терминологии и используют первые пришедшие в голову слова, лишь бы они были строго определены. В зрелом возрасте наступает обычно здравое отношение к терминологии. Старые же ученые, по мере того как их творческая активность угасает, начинают придавать терминологии все большее значение, пишут о ней, критикуют чужие работы не за их содержание, а за терминологию и т. д. Автор, по-видимому, уже перешел в третью категорию и поэтому не смог удержаться от включения этого параграфа в настоящую книгу.

Мы не будем говорить об общем положении с терминологией в области теории связи. Остановимся только на нескольких случаях, связанных с употреблением отдельных терминов и с отклонениями от общих языковых норм, с которыми приходится встречаться.

В свое время, рассматривая оптимальный прием сигналов на фоне небелого («окрашенного») гауссовского шума, В. А. Котельников ввел вспомогательный фильтр, преобразующий окрашенный шум в белый. Этот фильтр впоследствии стали называть «обеляющим», что вполне соответствует смыслу. Обелить — значит сделать белым. Но в последние годы все чаще этот фильтр называют «отбеливающим». Хорошо ли это? Слово «отбеливать» тоже означает сделать белым, но оно имеет явный химико-технологический оттенок.

Можно сказать «обелил свою совесть» и «отбеливает полотно», но наоборот вряд ли кто-либо скажет. Белый шум является все же математической абстракцией, и глагол «обелить» подходит к нему больше, нежели «отбелить».

От собственного имени Гаусс можно произвести два прилагательных — «гауссовский» и «гауссов», так же как от имени Петр можно произвести «петровские времена» и «петрово творенье». Однако все чаще и чаще, сначала в диссертациях, а теперь уже и в журнальных статьях пишут «гауссовый шум» и «гауссовая помеха». Но такое словотворчество в корне противоречит законам и духу русского языка. Никто ведь не скажет «петровый дворец». И никто, к счастью, не пишет «рэлеевые замирания» или «винеровый фильтр». Как же это могло произойти?

Лет десять назад такого слова встречать не приходилось. Писали и склоняли «гауссов шум», «гауссова шума», «гауссову шуму» и т. д. Но в выражении «канал с гауссовым шумом» форма «гауссовым» совпадает и с творительным падежом от «гауссовый». И как-то по инерции стали изменять и другие падежи и писать «гауссового шума» вместо «гауссова». Это было начало, а в последние 2—3 года стали уж прямо писать «гауссовый» даже в именительном падеже. Может быть я и ошибаюсь, но мне слышится в этом слове неуважение к памяти великого математика. Куда смотрят редакторы?

Как же лучше писать, гауссовский или гауссов? Мне кажется, что обе эти формы допустимы. Но все же некоторые нюансы в этих вариантах следует различать. Форма «гауссов» несколько архаична. Кроме того, она означает «принадлежащий Гауссу», и поэтому ее нежелательно применять к тем понятиям, с которыми Гаусс в свое время не сталкивался.

Форма «гауссовский» значит «относящийся к Гауссу» и имеет более широкую область применения. Так, по-видимому, можно писать «гауссово распределение», поскольку оно действительно получено Гауссом и принадлежит ему. Еще лучше избегать архаизма и писать «распределение Гаусса». Шумами же во времена Гаусса не занимались, и поэтому лучше применять форму «гауссовский шум», а не «гауссов шум» и не «шум Гаусса».

Часто встречаются технические и математические диссертации, написанные без соблюдения литературных и даже грамматических норм, бесцветные и неудобочитаемые. Замечания оппонентов по этому поводу считаются несущественными, и наоборот, если в отзыве указывается, что диссертация изложена хорошим языком и легко читается, то это воспри-

нимается не как выполнение обязательного условия, а как редкое достоинство.

В вопросах терминологии и словоупотребления следует внимательно прислушиваться к советам, даваемым филологами. Впрочем, здесь иногда тоже возникают курьезы и недоразумения. Так, несколько лет тому назад в «Литературной газете» была опубликована статья, в которой автор возмущался произвольными искажениями научных терминов и, в частности, отмечал, что раньше единицу информации называли «бит», а в последнее время все чаще пишут «байт». Ему и невдомек было, что «бит» и «байт» — это не разночтения, а две различные единицы.

Еще раньше, в той же «Литературной газете» в заметке о языке научной литературы, высказывалось недовольство словом «усреднять», которого якобы в литературном русском языке не существует, и предлагалось пользоваться «правильным» словом «осреднять». До сих пор не могут понять, чем префикс «у» хуже «о». Ведь и тот, и другой широко применяются для придания глаголу смысла законченного действия, хотя и имеют некоторые семантические различия. Например, мы говорим и пишем «остановить» и «установить», «окрасить» и «украсить», «охватить» и «ухватить», «оправлять» и «управлять» и т. д. Никто ведь не предложит вместо «удлинять» писать «одлинять». Чем же вызвало немилость привычное и ясное понятие «усреднить»?

К сожалению, при обучении инженеров мало внимания уделяют привитию им культуры речи. А это совершенно необходимо. Без точности языка не может быть ясности мысли

15.6. Несколько курьезов

В первые послевоенные годы, задолго до начала космической эры, одной из важнейших задач в области телевидения было отыскание способа передачи телевизионных программ на большие расстояния. Радиорелейных линий было очень мало, коаксиального кабеля еще меньше, а об искусственных спутниках Земли мечтали только писатели-фантасты.

В эти годы одновременно во многих странах возникла мысль использовать Луну в качестве пассивного ретранслятора. В частности, расчетами такой системы занимались и в лаборатории, которой в то время руководил автор. И вот на отчет, посвященный определению энергетических параметров и эффективности такой системы, поступила рецензия, в которой, в частности, говорилось:

«В отчете утверждается, что отражения от Луны позволят обеспечить передачу программ центрального телевидения в удаленные пункты страны примерно в течение 9—10 часов в сутки. Но там не учитывается, что каждый месяц наступает новолуние и в течение нескольких суток Луны на небе вообще нет, так что регулярного телевизионного вещания осуществить не удастся». Комментарии, надеюсь, излишни.

В одной рецензии на диссертацию содержалось примерно такое замечание: «Диссертант ошибочно утверждает, что величина A пропорциональна x^2. Как видно из его формул, A пропорциональна величине $x^2/2$».

Вообще терминам «пропорционально» и «обратно пропорционально» очень не везет. Много раз приходилось встречаться с утверждениями, что A обратно пропорционально B, тогда как в действительности имеет место лишь монотонное убывание A с возрастанием B. Такая неаккуратность в обращении с понятием пропорциональности вызывает часто забавные, а иногда и печальные последствия. Так, например, известно, что в канале с рэлеевскими замираниями вероятность ошибки приблизительно обратно пропорциональна мощности сигнала. В рукописи одного учебного пособия я вместо этого прочел следующее: «В канале с рэлеевскими замираниями вероятность правильного приема пропорциональна мощности сигнала». Автор пособия, как выяснилось в беседе с ним, считал обе формулировки совершенно эквивалентными. Тогда я его спросил, какой станет вероятность правильного приема при мощности сигнала $2P$, если при мощности P она равна 0,6. И только тут он призадумался.

15.7. Редакторы и корректоры

Что бы мы без них делали? Трудно переоценить роль редакторов и корректоров при выпуске книги и даже журнальной статьи. Часто мы говорим, что книга написана хорошим языком и легко читается, даже не подозревая, что рукопись была написана косноязычно и только упорная работа редактора позволила придать ей удобочитаемый вид. Мне посчастливилось чаще всего работать с хорошими редакторами. Но нет правила без исключения и о них-то я и хочу здесь рассказать.

В давние времена, когда я еще только начинал преподавать, а вопросы теории связи только начинали вводиться в программы обучения связистов, я написал и издал на ротаторе учебное пособие «Системы связи и помехозащищенность», в котором излагались для будущих инженеров основные положения теории вероятностей и помехоустойчивости. И вот редактор всюду изменил в рукописи термин «биномиальный» на «биноминальный». Когда я возмущенный прибежал к нему, он показал мне словарь, кажется, подготовленный Академией наук, в котором действительно слова «биномиальный» не было, а было слово «биноминальный». Как это могло произойти — я до сих пор не понимаю. Ведь этимологически слово «биноминальный» должно означать «имеющий два номинала», а вовсе не «относящийся к биному». Конечно, сейчас я не потерпел бы такой вольности и настоял бы на своем. Но тогда я был начинающим автором и бороться с редактором мне было не под силу. Так и вышло это пособие с «биноминальным распределением».

Несколько позднее, когда я издавал книгу в солидном издательстве, я обнаружил только в верстке, что «рэлеевские замирания» всюду изменены на «релеевские». В ответ на мой недоуменный вопрос редактор сослался на орфографический словарь, где указывалось, что слово «реле» пишется через «е», а не «э», и был очень удивлен, когда я заметил ему, что «рэлеевские замирания» происходят не от слова «реле», а от фамилии «Рэлей». Это написание так и сохранилось в книге, так как исправление всего набора сильно задержало бы ее выход в свет, а транскрипция иностранных фамилий в конце концов допускает некоторый произвол.

В другом издательстве в одной переводной книге редактор заменил выражение «М-ичный» на «М-й», что, конечно, затруднило понимание. Это выяснилось тоже только в верстке, а исправление потребовало бы переверстки значительной части набора. Пришлось согласиться с вариантом редактора и ограничиться разъяснением сокращения в подстрочном примечании.

Иногда самостоятельным «творчеством» занимаются и корректоры. Однажды в журнальной заметке корректор, не проконсультировавшись с редактором, заменил обозначение l. i. m. (предел в среднеквадратичном смысле) на lim (обычный предел), полагая, что точки поставлены ошибочно. В другой раз, в книге было использовано обозначение квантора общности: $\forall\ \varepsilon > 0$ (т. е. для всех ε, бо́льших нуля). Все было набрано верно, благополучно прошло верстку и сверку. В последний же момент, перед выпуском тиража, типографский

корректор спохватился, заметив, что буква А перевернута и поторопился «исправить» это, не спросив у редактора.

Еще хуже было с другой книгой, к которой автор, правда, отношения не имел. Типографский корректор обнаружил в последний момент на титульном листе непонятное ему выражение «транзисторные усилители», счел его опечаткой и заменил (без ведома редакции) на ясное и понятное любому пассажиру — «транзитные усилители».

К счастью, такие казусы происходят теперь редко. Наши высококвалифицированные редакторы, корректоры и все работники издательств и типографий делают все для того, чтобы наши книги и журналы были хороши во всех отношениях, успешно ликвидируют недостатки рукописи и заслуживают самой большой благодарности от авторов. Конечно, это относится в полной мере и к этой книге. За все те недостатки, которые в ней сохранились, полная ответственность лежит на самом авторе, который хочет ее закончить словами признательности, обращенными ко всем, помогавшим ему в ее создании.

Список литературы

1. **Агеев Д. В.** Активная полоса частотного спектра функции времени. — «Труды ГПИ», 1955, т. 11, № 1, с. 5—10.
2. **Attaway N. G.** Suppresed Sideband A-M Telephony. QST, 1974, v. 58, N 4, p. 49—50.
3. **Brennan D. G.** On the Maximum Signal-to-Noise Ratio Realizable from Several Noisy Signals.—«PIRE», 1955, v. 43, N 10, p. 1530.
4. **Бунимович В. И.** Флуктуационные процессы в радиоприемных устройствах. М., «Советское радио», 1951. 360 с.
5. **Вакман Д. Е.** Асимптотические методы в линейной радиотехнике. М., «Советское радио», 1962. 247 с.
6. **Вакман Д. Е.** Об определении понятий амплитуды, фазы и мгновенной частоты сигнала. — «Радиотехника и электроника», 1972, т. 17, № 5, с. 972—978.
7. **Введенский Б. А., Аренберг А. Г.** Радиоволноводы. М., Гостехиздат, 1946. 191 с.
8. **Винницкий А. С.** Модулированные фильтры и следящий прием ЧМ сигналов. М., «Советское радио», 1969. 547 с.
9. **Gabor D.** Theory of Communication.—«Journ. I. E. E.», 1946, v. 93, N 26, part III, p. 429—457.
10. **Галлагер Р.** Теория информации и надежная связь. Пер. с англ. под ред. М. С. Пинскера и Б. С. Цыбакова. М., «Советское радио», 1974. 719 с.
11. **Гоноровский И. С.** Радиотехнические цепи и сигналы. Ч. 1. М., «Советское радио», 1966. 439 с.
12. **Гоноровский И. С.** Радиотехнические цепи и сигналы. М., «Советское радио», 1971. 671 с.
13. **Горелик Г. С.** Колебания и волны. М.—Л., ГИТТЛ, 1950. 551 с.
14. **Железнов Н. А.** О принципиальных вопросах теории сигналов и задачах ее дальнейшего развития на основе новой стохастической модели. — «Радиотехника», 1957, т. 12, № 11, с. 3—12.
15. **Зюко А. Г.** Помехоустойчивость и эффективность систем связи. М., «Связь», 1972. 359 с.

16. **Зюко А. Г., Коробов Ю. Ф.** Теория передачи сигналов. М., «Связь», 1972. 280 с.
17. **Carson J. R.** Frequency Modulation.—«PIRE», 1922, v. 10, N 57, p. 243.
18. **Кловский Д. Д.** Потенциальная помехоустойчивость при разнесенном приеме дискретной информации. — «Радиотехника», 1961, т. 16, № 3, с. 22—30.
19. **Кловский Д. Д.** Теория передачи сигналов. М., «Связь», 1973. 376 с.
20. **Колмогоров А. Н.** Три подхода к определению понятия «количество информации» .— «Проблемы передачи информации», 1965, т. 1, вып. 1, с. 3—11.
21. **Коржик В. И.** Огибающая сигнала и некоторые ее свойства. — «Радиотехника», 1968, т. 23, № 4, с. 1—6.
22. **Коржик В. И.** Расширенное преобразование Гильберта и его применение в теории сигналов. — «Проблемы передачи информации», 1969, т. 5, вып. 4, с. 3—18.
23. **Котельников В. А.** Теория потенциальной помехоустойчивости. М.—Л., Госэнергоиздат, 1956. 151 с.
24. **Крамер Г., Лидбеттер М.** Стационарные случайные процессы. Пер. с англ. под ред. Ю. К. Беляева. М., «Мир», 1969. 398 с.
25. **Кушнир В. Ф., Ферсман Б. А.** Теория нелинейных электрических цепей. М., «Связь», 1974. 383 с.
26. **Левин Б. Р.** Теоретические основы статистической радиотехники. Ч. 1. М., «Советское радио», 1974. 550 с.
27. **Левитин Л. Б.** Перенос информации в идеальном фотонном канале. — «Проблемы передачи информации», 1965, т. 1, вып. 3, с. 71—80.
28. **Литлвуд Дж.** Математическая смесь. Пер. с англ. под ред. В. И. Левина. М., «Наука», 1973. 142 с.
29. **Манаев. Е. И.** О ширине полосы при приеме ЧМ сигналов. — «Радиотехника», 1948, т. 3, № 5, с. 54—61.
30. **Мидлтон Д.** Введение в статистическую теорию связи. Пер. с англ. под ред. Б. Р. Левина. М., «Советское радио», 1961, т. 1. 782 с.
31. **Nyquist H.** Certain Topics in Telegraph Transmission Theory.—«Transact. Am. I. E. E.», 1928, v. 47, N 2, p. 617—644.
32. **Винер Н., Пэли Р.** Преобразование Фурье в комплексной области. Пер. с англ. Ф. В. Широкова. М., «Наука», 1964. 267 с.
33. **Петрович Н. Т.** Передача дискретной информации в каналах с фазовой манипуляцией. М., «Советское радио», 1965. 263 с.

34. **Петрович Н. Т.** «Кто Вы?». М., «Молодая гвардия», 1970. 239 с.
35. **Рытов С. М.** Модулированные колебания и волны. — «Труды ФИАН», 1940, т. II, вып. 1, с. 41—133.
36. **Рытов С. М.** О некоторых «парадоксах», связанных со спектральным разложением. — «Успехи физических наук», 1946, т. 29, вып. 1—2, с. 147—160.
37. **Смирнов В. А.** Основы радиосвязи на ультракоротких волнах. М., Связьиздат, 1957. 819 с.
38. **Стейн С., Джонс Дж.** Принципы современной теории связи и их применение к передаче дискретных сообщений. Пер. с англ. под ред. Л. М. Финка. М., «Связь», 1971. 376 с.
39. **Стратонович Р. Л.** Избранные вопросы теории флюктуаций в радиотехнике. М., «Советское радио», 1961. 558 с.
40. **Тихонов В. И.** Один способ определения огибающей квазигармонических флюктуаций. — «Радиотехника и электроника», 1957, т. 2, № 4, с. 562—568.
41. **Тихонов В. И.** Статистическая радиотехника. М., «Советское радио», 1966. 680 с.
42. **Тихонов В. И., Кульман Н. К.** Нелинейная фильтрация и квазикогерентный прием сигналов. М., «Советское радио», 1975. 704 с.
43. **Трахтман А. М.** Введение в обобщенную спектральную теорию сигналов. М., «Советское радио», 1792. 352 с.
44. **Турин Дж.** Лекции о цифровой связи. Пер. с англ. под ред. Р. З. Хасьминского. М., «Мир», 1972. 104 с.
45. **Финк Л. М.** Соотношения между спектром и мгновенной частотой сигнала. — «Проблемы передачи информации», 1966, т .2, вып. 4, с. 26—38.
46. **Финк Л. М.** Теория передачи дискретных сообщений. М., «Советское радио», 1970. 727 с.
47. **Fleming A., Fortescue C. L. and oth.** Notes on Modulation. —«Nature», 1930, v. 125, p. 92, 198, 271, 306.
48. **Френкс Л.** Теория сигналов. Пер. с англ. под ред. Д. Е. Вакмана. М., «Советское радио», 1974. 343 с.
49. **Харкевич А. А.** Спектры и анализ. М.—Л., ГИТТЛ, 1952, 192 с.
50. **Harman W. W.** Instantaneous Frequency.—«PIRE», v. 42, N 3, p. 599.
51. **Хворостенко Н. П.** Статистическая теория демодуляции дискретных сигналов. М., «Связь», 1968. 335 с.
52. **Hok G.** Frequency Modulation and Instantaneous Frequency,—«PIRE», 1953, v. 41, N 12, p. 1786.

53. **Хургин Я. И., Яковлев В. П.** Финитные функции в физике и технике. М., «Наука», 1971. 408 с.
54. **Shekel G.** «Instantaneous» Frequensy. — «PIRE», 1953, v. 41, N 4, p. 548.
55. **Shekel G.** On the Term Instantaneous Frequency.— «PIRE», 1954, v. 42, N 6, p. 1024.
56. **Шеннон К.** Математическая теория связи. — В кн.: Работы по теории информации и кибернетике. Пер. с англ. под ред. Р. Л. Добрушина и О. Б. Лупанова. М., «ИЛ», 1963, с. 243—332.
57. **Шеннон К.** Теория связи в секретных системах. — В кн.: Работы по теории информации и кибернетике. Пер. с англ. под ред. Р. Л. Добрушина и О. Б. Лупанова. М., «ИЛ», 1963, с. 333—402.
58. **Шеннон К.** Предсказание и энтропия печатного английского текста. — В кн.: Работы по теории информации и кибернетике. Пер. с англ. под ред. Р. Л. Добрушина и О. Б. Лупанова. М., «ИЛ», 1963, с. 669—686.
59. **Шеннон К.** Бандвагон. В кн.: Работы по теории информации и кибернетике. Пер. с англ. под ред. Р. Л. Добрушина и О. Б. Лупанова. М., «ИЛ», 1963, с. 667—668.
60. **Шкловский И. С.** Вселенная, жизнь, разум. М., «Наука», 1976. 336 с.
61. **Шрейдер Ю. А.** О семантических аспектах теории информации. — В кн.: Информация и кибернетика. Под ред. А. И. Берга. М., «Советское радио», 1967, с. 15—47.

ОГЛАВЛЕНИЕ

Стр.

Предисловие 3

1. Об ошибках, парадоксах и задачах этой книги (Вместо введения) 4

1.1. Об этой книге 4
1.2. Ошибки 7
1.3. Парадоксы 9

2. Спектр и мгновенная частота

2.1. Спор о спектре амплитудномодулированного сигнала 16
2.2. Спор об узкополосной частотной модуляции . . 21
2.3. В чем корень ошибки? 24
2.4. Обжегшись на молоке 29
2.5. Попробуем быть последовательными . . . 33
2.6. К теории ЧМ альтиметра 36
2.7. Существуют ли реально спектральные составляющие? 37
2.8. Ошибки живучи 39

3. О комплексном представлении сигнала

3.1. Обобщение символического метода 41
3.2. Однозначно ли определены огибающая и мгновенная частота? 43
3.3. Что такое преобразование Гильберта? . . . 46
3.4. Другие определения сопряженного сигнала . . 50
3.5. Конкурс определений 53
3.6. Как же быть с финитными сигналами? . . . 56
3.7. Занимательная задача 59

4. Как не следует пояснять теорему Котельникова

4.1. Сущность теоремы 61
4.2. Попытки наглядного пояснения теоремы . . . 64
4.3. Ослабленная теорема Агеева 66

Стр.

4.4. Доказательство теоремы Агеева 66
4.5. Можно ли передать мегабит за секунду в полосе 1 Гц? 69

5. Предел или барьер Найквиста 72

6. Еще о спектре сигнала и его мгновенной частоте

6.1. Сущность вопроса 87
6.2. Некоторые заблуждения 88
6.3. Поиски 91
6.4. Находки 96
6.5. Разочарование 100
6.6. Кое-что о линейных цепях 100

7. Вокруг частотной модуляции

7.1. Применяется ли частотная модуляция на практике? 104
7.2. Ширина спектра при угловой модуляции . . . 109

8. На одной боковой

8.1. Рекорды однополосной модуляции 115
8.2. Как записать однополосный сигнал? 119
8.3. Как выглядит однополосный сигнал? 120
8.4. «Формула Костаса» 123
8.5. Как детектировать однополосный сигнал? . . . 130
8.6. Фантазии о квадратичном детектировании ОМ сигналов 140
8.7. О требуемой точности восстановления несущей частоты 144

9. Связь с собратьями по разуму

9.1. Готовьтесь к межпланетной связи 150
9.2. Передача в диапазоне сантиметровых волн . . 151
9.3. Связь в оптическом диапазоне 154
9.4. Где выход? 158

10. Лучше наилучшего

10.1. Изобретатели «вечного двигателя» 161
10.2. Широкополосная частотная манипуляция . . . 166
10.3. Интервальная манипуляция 170
10.4. «Сверхоптимальный» прием 177
10.5. Давайте разберемся 185

11. Парадокс импульсной модуляции

11.1. Порог помехоустойчивости 195
11.2. Пропускная способность 199
11.3. Небольшие уточнения 200

12. Берегитесь неточных формулировок

12.1. Занимательные парадоксы 207
12.2. Количество информации, передаваемой в гауссовском канале с неопределенной фазой 208
12.3. О рэлеевских замираниях 212

13. Как оценивать верность передачи сообщений?

13.1. Вероятность ошибочного декодирования . . 217
13.2. Эквивалентная вероятность ошибки 219
13.3. Где применять разнесенный прием? 225
13.4. Критерии верности телефонного сигнала . . . 229
13.5. Чем кормить кошку? 234

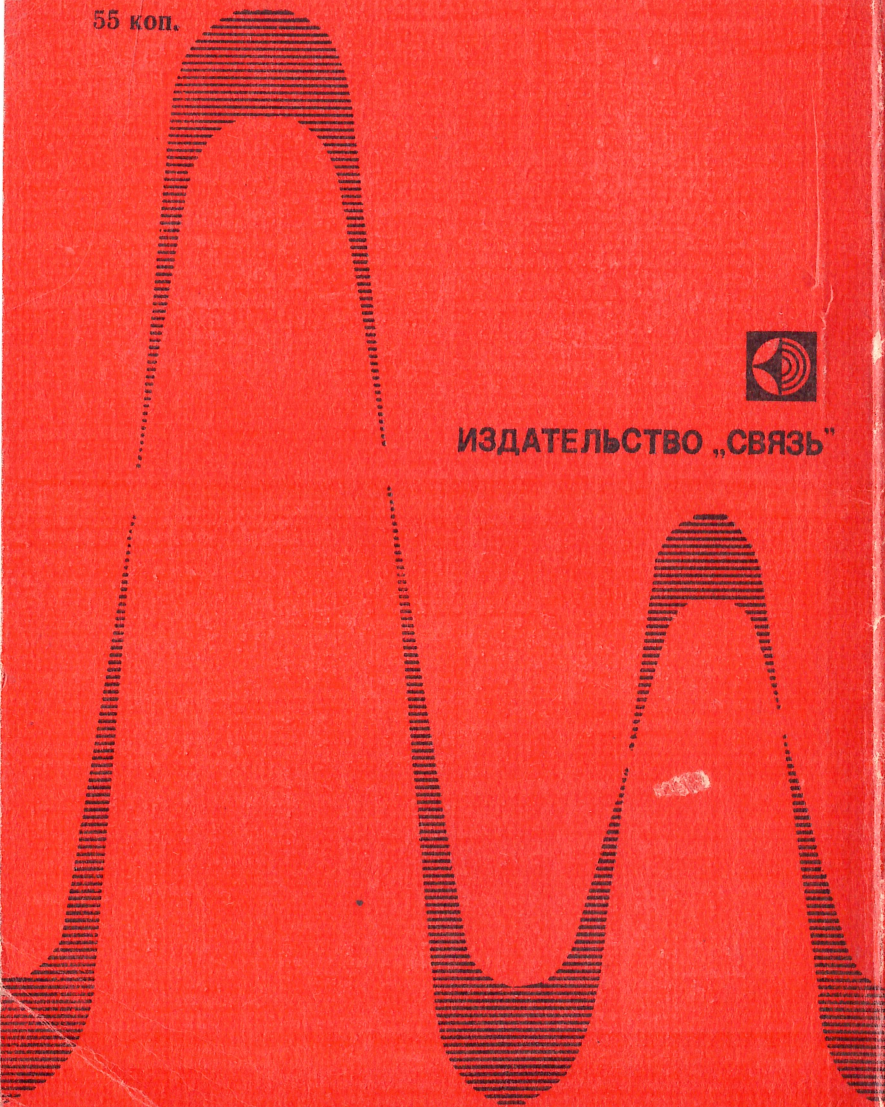
14. Немного информации

14.1. О постулатах теории информации 236
14.2. Первые три условия 239
14.3. Информация и время 243
14.4. Семантическая информация 245

15. Мелочи

15.1. ФМ = ДБП? 251
15.2. Немного о белом шуме 252
15.3. Почему сигналы случайны? 255
15.4. Оценка энтропии текста методом отгадывания . 256
15.5. Кое-что о терминах 260
15.6. Несколько курьезов 262
15.7. Редакторы и корректоры 263
Список литературы 266

55 коп.



ИЗДАТЕЛЬСТВО „СВЯЗЬ"

А. М. ФИНК СИГНАЛЫ, ПОМЕХИ, ОШИБКИ...